WHO'S AFRAID OF FILE FORMAT OBSOLESCENCE? EVALUATING FILE FORMAT ENDANGERMENT LEVELS AND FACTORS FOR THE CREATION OF A FILE FORMAT ENDANGERMENT INDEX

Heather Ryan

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Information and Library Science.

Chapel Hill
2014

Approved by:

Christopher A. Lee

Kurt Bollacker

Diane Kelly

Richard Marciano

Helen Tibbo

ABSTRACT

Heather Ryan: Who's Afraid of File Format Obsolescence Evaluating File Format Endangerment Levels and Factors for the Creation of a File Format Endangerment Index
(Under the direction of Christopher A. Lee)

Much digital preservation research has been built on the assumption that file format obsolescence poses a great risk to the continued access of digital content. In an endeavor to address this risk, a number of researchers created lists of factors that could be used to assess risks associated with digital file formats. This research examines these assumptions about file format obsolescence and file format evaluation factors with the aim of creating a simplified file format endangerment index.

This study examines file format risk under a new lens of file format endangerment, or the possibility that information stored in a particular file format will not be interpretable or renderable in human accessible means within a certain timeframe. Using the Delphi method in two separate studies, this exploratory research collected expert opinion on file format endangerment levels of 50 test file formats; and collected expert opinion on relevance of 21 factors as causes of file format endangerment.

Experts expressed the belief that generally, digital information encoded in the rated file formats will be accessible for 20 years or more. This indicates that file format experts believe that there is not a great deal of short-term risk associated with encoding information in the rated file formats, though this does not preclude continued engagement with preservation activities for these and other file formats. Furthermore, the findings show that only three of the dozens of file format evaluation factors discussed in the literature exceeded an emergent threshold level as causes of file format endangerment: *rendering software available*, *specifications available*, and *community/3rd party support*.

These factors are ideal candidates for use in a file format endangerment index. Such an index allows only for the inclusion of formative indicators, or factors that indicate a cause of file format endangerment. In contrast, a scale can contain factors that reflect, rather than indicate a cause of a particular phenomenon. The three factors shown to be the most relevant as causal indicators of file format endangerment, *rendering software available*, *specifications available*, and *community/3rd party support* are the best candidate indicators to build into the index. The intention is to construct and validate an index using these three candidate factors as part of a future research agenda.

For Joe.

ACKNOWLEGEMENTS

It's been a long, wild ride. It's been a ride that has never once been lonely. Many, many people have been there with me on the path leading up to and through this moment, and I'd like to express my tremendous gratitude to all of you. I'm going to try to start at the beginning and catch everyone who has helped and encouraged me along the way.

I am going to follow in my committee chair's footsteps a little here and thank the most important person first: Joseph Dennis Ryan. I don't know where to start except to say that he has been the most incredible partner imaginable. His unwavering strength, resilience, support, patience, kindness, love, compassion, humor, prodding, encouragement, and well, everything he's done for me in these past years have helped me get to where I am now. I honestly don't know if I could have done it without him.

I had many wonderful teachers in high school. Three have stayed with me over all of these years. First, was Dr. B, or Dr. Bishop. She asked a lot of her students, but was always kind and exceptionally supportive. It is in part because of her that I have such a love of literature, writing, and critical thinking today. She passed away the year after I took her class, but I am still influenced by her joy and passion for knowledge and teaching. The second is Raul Dorn. There was some special kind of magic that happened in his art classes that I've never known since. We were a tight-knit group of misfits back then, but this smart, kooky guy made us feel like we were part of something, like we belonged, like we could make wonderful things happen. The third is Jim Smith. He taught AP History, my first seminar class, where he asked us to really think about and question everything around us. He also exhibited a true joy for teaching that serves as a model for my teaching today.

In my years as an undergrad student at New Mexico State University, one professor shines the brightest as having a profound effect on my belief in myself and on the direction my life took after meeting him: Dr. Alvin Keaton. I have to speak somewhat more at length about him now; first because he made such a difference in my life, and second because he passed away just a few weeks ago. One of the first things he taught me was that the best way to learn something is to teach it. In his philosophy class, he excused the students from taking all future exams who made the top grades on his first exam, with the stipulation that they would serve as exam tutors for the rest of the semester. I was one of these “lucky” tutors, but of course, I had to study ten times harder to be a tutor than I ever did to take exams. But this was just the beginning of my relationship with Dr. Keaton. I used to meet with him and my then boyfriend, Cassidy King, for breakfast every Friday. He would buy us breakfast and we would talk endlessly about philosophy, science, and life. At these breakfasts he would often look at me and, shaking his head, ask me why I wasn’t achieving more, why I wasn’t living up to my potential. Before Al Keaton, I can honestly say that I didn’t believe I had much potential beyond working menial jobs and being a “starving artist.” I knew I was wildly curious about the world and was constantly thinking about how everything worked and fit together; but I had never had an outlet for these thoughts, and I had never received validation from the external world that the things I think and care about were actually worth a damn. It was from those moments with Al that the first sparks were ignited that propelled me to where I am today. The last time I had any real interaction with him, I was headed off to Houston, my first time living away from my hometown. He gave me $20 and told me to read Kevin

Kelly's *Out of Control*, both of which had a big impact on me.

As soon as I got to Houston, I found and devoured a copy of *Out of Control.* It left me with a sense that I wasn't quite as alone in my thinking as I thought. After living in a desert, both literally and figuratively, it was like a long, cool drink of water. Years after my last encounter with Al, I found myself exploring digital preservation as a masters student at the University of Denver. In my studies, I stumbled across *Time & Bits: Managing Digital Continuity*, a book about a collaborative conversation between members of the Getty Conservation Institute, the Getty Information Institute, and the Long Now Foundation. I read the book, and lo and behold, there was Kevin Kelly again. And there was Long Now, and they were accepting members. As soon as I could scrounge together enough money for the membership dues, I joined. Not too long after, I flew out to San Francisco for a Brian Eno art exhibit they were hosting, where I met and talked to Kevin Kelly about my research. After a short conversation with him and Alexander Rose, we agreed that I would do my master's practicum with them that following summer. It's a bit of a long story, but suffice it to say that the people of Long Now have had a big impact on my research. In particular, Alexander Rose, Laura Welcher, and Kurt Bollacker (one of my committee members, whom I will talk more about later) have all been tremendously gracious and supportive of me. I would also like to acknowledge my long-time friend, Richard Mortimer Humphrey, for supporting my goals and hosting me in San Francisco while I worked with Long Now.

During my master's program at the University of Denver, I had two advisors who supported me in my digital preservation research: Dr. Denise Anthony and Dr. Rich Gazan. It was in Rich's classes that I first realized that I wanted to get my Ph.D. Denise has been my unfailing advocate from when she was first my advisor to today.

While many of my friends were not directly involved with my research, they were all wonderful to and supportive of me as I worked toward my goal of getting my Ph.D. First, I'd like to mention one of my oldest friends, Rowan Russell. In Colorado, the most notable of my support-team friends were: Radovan Anzulovic, Matthew Brown, Andrea Davis, Nancy Dituri, Michael Doherty, Donna Goben, Hayley Helmericks, Jaime Jennings, Tara Leedom, Mark Phillips, Julie Rockwell, Amanda Rybin, Doug Spencer, Carissa Sullivan, Katie Thomas, and Zarin Xavier. Of these people, I would like to give extra acknowledgement to Donna Goben, my new honorary sister. In North Carolina, my most notable supporters were: Earl Bailey, Mike Brown, Michelle Corbin, Josh Hockensmith, Erin Kawamata, Kelly Marks, Renee McBride, Margarite Nathe, Ericka Patillo, Rachel Powers, Terrell Russell, Mike Sonnichsen, Liz Wing, and Kam Woods. Of these people I would like to particularly thank Margarite Nathe and Josh Hockensmith for boosting my spirits every week at game night, camping and beach trips together, and for being the best damn wedding helpers a lady could ask for. I'd also like to thank Earl Bailey and Michelle Corbin for introducing us to the fantastic world of "Euro" games, all of the shared meals, and for being family. I'd also like to thank Mike Brown for all of the support and encouragement he gave me during his brief stint as my writing coach. I'd also like to thank all of my friends on Facebook from all over the world who cheered me on when I most needed it. Last, but not least of my friends, I'd like to acknowledge those doctoral student colleagues not already mentioned here. It's been a long, wild ride for all of us. I've learned a little something from all of them and I am grateful for it. I wish them all the best and hope our paths cross again, preferably in less challenging circumstances.

A great number of people at the University of North Carolina at Chapel Hill School of

Information and Library Science helped me in small and big ways throughout my career as a doctoral student. To name a few of the faculty: Dr. Gary Marchionini, Dr. Barbara Wildemuth, Dr. Evelyn Daniel, Dr. Reagan Moore, Dr. Jeff Pomerantz, Dr. Deborah Barreau, and Paul Jones. Staff to whom I am grateful are: Lara Bailey, Aaron Brubaker, Tammy Cox, Wanda Monroe, Kaitlyn Murphy, Michael Penny, Susan Sylvester, and Shaundria Williams. I'd like to note especially that it was Deb Barreau who talked me into applying for the program. During the time that I was able to spend with her, she was unfailingly kind and compassionate. Though I did not spend time with her regularly, her passing affected me deeply, and I miss her kindness and support.

I'd also like to acknowledge my new colleagues at the University of Denver: Mary Stansbury, Clara Sitter, Shimelis Assefa, and Krystyna Matusiak. Joining their team has been perhaps the biggest motivator to finish my dissertation! Thank you! I'd also like to extend special thanks to Jen LaBabera for her last minute help with tables.

Since I started the program in 2008, several people in my life, not mentioned yet, passed away. Three of these people were particularly important to me in different stages of my life. I have to admit that it is particularly difficult to relive the losses, but I believe it's important to honor their memory here for posterity. First is tS. Pureé Tomatoes, the backwards saint. Describing Tomatoes and his importance to me would fill a book on its own. He was one of the most brilliant and genuine people I've ever known. There never was and never will be another like him. Second is Dylan Williams. He was the first person I ever interviewed in my zine-making days. He was a brilliant comics artist and started a great comics publishing empire, Sparkplug Comics. We never met in person, but we stayed in touch over the past twenty years. When I entered the Ph.D. program, he was full of kindness and encouragement. His passing grieved me terribly. Last, and most importantly, is my Grandma, Barbara Bowden. Of all the people in my blood family, she supported me the most. She was the great matriarch of my family, and while she was not always the kindest, she was the strongest. I always think of her when I have something particularly difficult to do. She took on all of life's challenges with a no-nonsense, single-minded fire, and she always succeeded in her goals. I owe a lot of my successes in life to her. Look Grandma, I'm finishing my Ph.D.!

It is with mixed emotions that I mention my blood family here. Obviously, I would not be here if it were not for them. Their confused and fearful encouragement certainly helped me to a degree. I do love them and recognize that they all gave me the most that they could with what they had.

Where I was a misunderstood and frightening black sheep in one family, I found myself right at home in another. Thank you to Dr. Dennis Ryan, Linda Ryan, Dr. Jennifer Ryan, Dr. Tim Bryant, and Miranda Bryant (and special thoughts for Grace Bryant) for welcoming me into your family. Thank you for all of the books (*books*!!), the understanding, the support, and the love. It means more to me than you can imagine. And I am thrilled to finally become a member of the illustrious the Dr. (B)Ryan(t) club!!

I want to give some very, very special thanks to another part of my non-traditional family, the Schroecks. Not only did they provide a home, love, care, a sister (Anna!!) and support for my son, but also they served as a model of what I might be able to achieve in my life. Their kindness, stability, and successes in life served as a solid guidepost to what I wanted to accomplish in mine. I would like to acknowledge Anna: she has never ceased to impress me with her intelligence, strength, kindness, mischievousness, patience, poise, and beauty. She is growing into an amazing young woman, and it is an inspiration to see. I would also like to acknowledge my son, Adrian. The love I feel for him surpasses any feeling I've ever had. It is the most beautiful gift. I am so proud of him and I have thrilled at watching him grow in the brilliant, curious, focused, hard working, playful, handsome young man he is today.

Getting down to the nitty gritty of my actual dissertation, I'd like to thank my intrepid pilot study participants. Their hard work and feedback were tremendously helpful. Their answers were so thoughtful that I regretted the fact that they would not be included in the data analysis. The pilot participants were: Jefferson Bailey, Shane Beers, Brendan Donahe, Chien-Yi Hou, Marcos Martinez, Matthew Mayernik, Trevor Owens, Sam Meister, Dave Pcolar, Terrell Russell, Bill Schulz, and Kam Woods.

I'd like to extend a very special thanks to all those who participated in the study. I am overwhelmed with gratitude for the amount of time and depth of thought they put into this study. They made an invaluable contribution to my research, and I am truly indebted to all of them. Thank you very much to: Stephen Abrams, Micah Altman, Kevin Ashley, Euan Cochrane, Maurice deRooij, Kevin DeVorsey, Mark Evans, Carl Fleischhauer, Jay Gattuso, Andrea Goethals, Sergiu Gordea, Hans Hofman, Matt Holden, Andrew Jackson, Catherine Jones, Steve Knight, Peter May, Jerome McDonough, Erin O'Meara, Nicholas Taylor, and William Underwood.

I feel incredible gratitude toward my dissertation committee. I'll start with Dr. Kurt Bollacker, my external committee member. As I mentioned earlier, I met him during my practicum experience with the Long Now Foundation. Though everyone at Long Now was kind, enthusiastic, and supportive, Kurt was especially so. He quite unexpectedly took me under his wing and provided me with invaluable mentorship. After I was accepted to the Ph.D. program, I immediately asked him to be on my dissertation committee. I am thrilled that he accepted. The kind and enthusiastic mentorship he exhibited during my master's practicum has not ceased. It is with a cringing smile that I share what he wrote on my practicum evaluation: "She's too polite!" I'm still working on that.

Dr. Richard Marciano and I arrived at UNC at the same time and I started working with him almost immediately. Richard was always kind, funny, enthusiastic, and encouraging. One of his great skills is that of a connector. He connected me with a number of great people and great resources throughout my career as a doctoral student. Both he and Marc Fresko independently helped me formulate the notion that file formats can be viewed similarly to living organisms.

Dr. Diane Kelly is special to me for several reasons. First, she provided invaluable direction and advice in the development of my research methods. Second, her intelligence, wit, and genuine love for her work and this field have been truly inspiring. I see her as a great model researcher and educator. If I manage to be anything like her in my career, I will consider it a great success. Third, it was in her Research Methods class that I met my husband, Joe. While all of her contributions to my research and career are great, this special contribution to my life is the best. Thank you, Diane, for everything.

Dr. Helen Tibbo will always, always have a special place in my academic and personal heart. Her trailblazing work in this field is astonishing. Without her foresight in brining digital curation education to information and library science programs at both the masters' and the doctoral level, I doubt I would be here at all. She brought me into the program, advised me, provided me with opportunities to travel, introduced me to important figures in the field, fought for me, and ceaselessly encouraged me to keep going to the end. I could not have asked for a better advisor, mentor, and committee member.

Similarly, Dr. Cal Lee has surpassed my expectations in a chair and advisor. I have joked about how thoroughly he reviews and edits my work, but in all honesty, it is his thoroughness and attention to accuracy and detail that have pushed me to grow and improve. I am very grateful to have a mentor and advisor who took the time to always push me to be and do better. He has provided me with many opportunities connect with noteworthy professionals in my research area. It is thanks to his support of me and my research that I have successfully launched my career in academia. He is also one of the busiest people I know, both as an academic and as a father and husband. Knowing this, it fills me with gratitude that he has given so much of his time in shepherding me to this point. His unceasing energy, care, innovation, and drive are an inspiration and model for me as I move forward in my career.

Last, but by no means least I would like to thank the Institute of Museum and Library Services (IMLS) and the National Science Foundation (NSF). It was thanks to an IMLS grant, *DigCCurr II* (#RE-05-08-0060-08), and an IMLS cooperative agreement award, *Closing the Digital Curation Gap II* (#LG-05-09-0040), that I received tuition and stipend fellowships for most of my time at UNC. I also received tuition and a stipend for one semester from NSF, (NSF III #1247471), for which I am very grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIHT	Archive Ingest and Handling Test
AIP	Archival Information Package
AONS	Automated Obsolescence Notification System
APSR	Australian Partnership for Sustainable Repositories
CRL	Center for Research Libraries
DIAS	Digital Information Archiving System
DIP	Dissemination Information Package
DPSP	Digital Preservation Software Platform
DRM	Digital Rights Management
DROID	Digital Record Object Identifier
DSTC	Cooperative Research Centre for Enterprise Distributed Systems Technology
EDVAC	Electronic Discrete Variable Automatic Computer
ENIAC	Electronic Numerical Integrator and Computer
FIDO	Format Identification for Digital Objects
FITS	File Information Toolset
FTC	Fundação para Ciência ea Tecnologia, Portugal
GDFR	Global Digital Format Registry
GPHIN	Global Public Health Intelligence Network

GTRI	Georgia Tech Research Institute
HTML	Hypertext Markup Language
HUL	Harvard University Library
INFORM	INvestigation of Formats based on Risk Management
IRB	Institutional Review Board
iRODS	integrated Rule Oriented Data System
IUCN	International Union for Conservation of Nature
JHOVE	JSTOR/Harvard Object Validation Environment
JSTOR	Journal Storage
NCAST	National Archives and Record Administration’s Center for Advanced Systems and Technologies
NDIIPP	National Digital Information Infrastructure & Preservation Program
OAIS	Open Archival Information System
OCLC	Online Computer Library Center
OPF	Open Planets Foundation
PANIC	Preservation Architecture for New Media and Interactive Collections
PRESERV	Preservation Eprint Service
PVA	Population Viability Analysis
ROAR	Registry of Open Access Repositories
SCAPE	SCALable Preservation Environments

SDB	Safety Deposit Box
SIP	Submission Information Package
SOA	Service Oriented Architecture
TNA	The United Kingdom National Archive
TRAC	Trustworthy Repositories Audit and Certification
UDFR	Unified Digital Format Registry
UN ISDR	United Nations International Strategy for Disaster Reduction
XML	Extensible Markup Language

CHAPTER 1 : INTRODUCTION

With the creation of the first computer program, there came the need for a method to save the output of the program as well as the program itself. Developers of one of the earliest computers created the Electronic Numerical Integrator and Computer (ENIAC), which produced content encoded in the first and only "file format" of its time. The ENIAC could store up to twenty numbers as pulses in electronic tubes. Its successor, the Electronic Discrete Variable Automatic Computer (EDVAC), came with a new method of storage, the mercury delay-line, which could store 1,000 bits of information. The mercury delay-line was, for all intents and purposes, the beginning of persistent read/write storage of computer programs and operational data (Campbell-Kelly and Aspray, 2004).

Since then, people have been using computers to create and collect an increasing amount of binary information. This capability has not only changed the way we preserve our cultural memory, but also our ability to compute and create new knowledge. In his 1984 report on *The Archival Appraisal of Machine-readable Records*, Harold Naugler commented that "technology has introduced a different type of information, namely 'processable information.' The information is accessible, interpretable, manipulable, and transmittable only by automated or electronic means" (1984, p. 14).

David Bearman said of this shift: "Over the next twenty-five years we can expect to take part in a worldwide effort to represent the entire corpus of civilization in digital form"

(1994b). Seamus Ross noted that digital information is "changing the way in which our culture is recorded…digital information is a cultural product" (2000, p. 3).

Preserving access to cultural products in digital form has posed many challenges. Information stored in digital form has a much shorter shelf life than information stored in analog forms. Digital information does not stand up to "benign neglect" (Bennett, 1997) nearly as well as information stored on analog media. Whereas information recorded on stone and paper can ideally last thousands of years, digital media such as magnetic and optical disks are predicted to last only decades (Bollacker, 2010).

Even if the bitstreams survive the decay of the physical storage media, they still require specific software and operating systems to translate them into human-usable form. The rate at which software changes makes it increasingly difficult to maintain meaningful, trustworthy access to digital information over time. One such challenge is associated with digital file formats, or the "internal structure and/or encoding of a file which allows it to be interpreted or rendered in human accessible form" (The National Archive [TNA], 2005, p.1).

File format obsolescence is a phrase commonly used to describe the phenomenon that occurs when information stored in a particular file format is no longer accessible using current technology. Although it has often been the focus of research and discussion in digital preservation throughout the years, there are surprisingly few formal definitions of file format obsolescence in the literature. Even a paper entitled *Defining File Format Obsolescence* (Pearson and Webb, 2008) does not present a formal definition of the term.

According to Merriam-Webster's Collegiate Dictionary, obsolescence is "the process of becoming obsolete or the condition of being nearly obsolete," where obsolete is defined as, "no longer useful or no longer in use" (Merriam-Webster, Inc., 1994, p. 803). This being the case, file format obsolescence could be defined as the process of a file format becoming no longer useful or no longer in use.

While the term *file format obsolescence* is still useful to describe a state in which a file format is no longer in use, I will use the term *file format endangerment* to describe the possibility that information stored in a particular file format will not be interpretable or renderable using standard methods within a certain timeframe. This term will be used in a way that is similar to its application to animal species. According to Merriam-Webster, *endanger* means, "to bring into danger or peril," where an endangered species is "a species threatened with extinction," or more broadly, "anyone or anything whose continued existence is threatened" (Merriam-Webster, 1994, p. 381). A file format is not threatened with extinction or a discontinued existence; rather the threat is to the user's ability to access information from a file that is encoded in that format.

Using the phrase *file format endangerment* provides a new perspective for studying the nature of these risks. By studying a file format's ability to be rendered as being similar to animal species endangerment, potentially useful parallels may be created that can lend new insight into the problem. Animal species have been studied for hundreds of years, and the methods used to document and assess the factors that contribute to their thriving or extinction can be applied to the viability or inaccessibility of the different "species" of file formats. From this we can learn which factors most heavily contribute to the risk of file format endangerment, and we can use this knowledge to identify this risk and take action to ameliorate it. Finally, the term "endangerment" embodies a sense of hope and urgency that hopefully incites action; much more so than the term obsolescence, which emits a sense of loss that is irreparable.

Researchers and practitioners have developed a number of theoretical frameworks, tools, and systems to address the social and technical challenges inherent in digital preservation. The Consultative Committee for Space Data Systems (CCSDS) created the Reference Model for an Open Archival Information System (OAIS) (2002; 2012), the Digital Curation Centre developed the Digital Curation Lifecycle model (Higgins, 2008), the Online Computer Library Center (OCLC) and the Center for Research Libraries (CRL) released the Trustworthy Repositories Audit and Certification (TRAC) Criteria and Checklist (2007; 2012); all of which provide frameworks from which the digital preservation community has created workflows, tools and systems that address the challenges of file formats in digital preservation.

The OAIS Reference Model and the Criteria and Checklist both indicate a need for technology monitoring to track potential risks associated with file format endangerment. The OAIS Reference Model has a Monitor Technology function that is “responsible for tracking emerging digital technologies, information standards and computing platforms (i.e. hardware and software) to identify technologies which could cause obsolescence in the Archive’s computing environment and prevent access to come of the archives current holdings” (CCSDS, 2012, p. 4-14). The functional entity B3.2 in the TRAC Criteria and Checklist states that a repository should have “mechanisms in place for monitoring and notification when Representation Information (including formats) approaches obsolescence or is no longer viable” (OCLC & CRL, 2007, p. 31).

A close examination of the existing digital preservation tools and systems, however, reveals a gap in this area of need: none of them have yet to operationalize file format risk monitoring. Many of the tools and systems discussed in the literature review indicate that their file format risk analysis components will come from PRONOM, a registry of file format information; but PRONOM does not currently contain information on file format risk information. Systematic file format endangerment monitoring is essential to the functioning of these systems and to the broader community's need to understand the risks that may be associated with the file formats in their digital collections.

While there have been a number of attempts to make systematic file format risk analysis possible, such as the numerous lists of file format evaluation criteria shown in the literature review, they have yet to be operationalized. This lack of file format risk information leaves the community with only a sense that file formats are a problem for digital preservation, but little information about the degree to which certain file formats are endangered. This lack of data makes it difficult for researchers and practitioners to make solid decisions about policies, models, tools, and systems that involve file formats.

The problem of file format risk analysis is similar to problems found in the research area of conservation biology. Conservation Biology uses a method called Population Viability Analysis (PVA) where data about individual species is collected and analyzed for several formative indicators, such as population count and environmental health that feed into the International Union for Conservation of Nature (IUCN) Red List Index for Threatened Species.

1.1. Research Framework

I have examined the research methods used in digital preservation and conservation biology and have used them to inform the design of a methodology that addresses the need for data-informed file format endangerment risk analysis. This methodology includes the following steps:

1. Collect baseline knowledge of file format endangerment levels
2. Select formative indicators for a file format endangerment index
3. Collect data for file format endangerment index indicators
4. Test and validate the file format endangerment index
5. Begin implementation of the file format endangerment index to create file format endangerment ratings that can be used to inform digital preservation decision-making

Merriam Webster defines an index as, “a number derived from a series of observations and used as an indicator or measure” (1994, p. 591). In the context of a file format endangerment metric, the derived number will be a calculated file format endangerment level, and the indicators are the factors that directly affect the endangerment level of a particular file format.

This research addressed portions of the first three of these steps, and through this research I sought to answer the following four research questions:

1. Do digital preservation experts believe that certain file formats pose a risk for digital preservation?
2. Which file formats do digital preservation experts believe are more endangered than others?
3. What are the most relevant formative indicators of file format endangerment, and how can these indicators be measured?
4. How effectively can the expert-chosen file format endangerment factors be applied to rating file format endangerment?

First, I collected baseline expert opinion of file format endangerment levels for fifty test file formats. I collected this information by employing the Delphi method during which I asked a group of digital preservation experts to rate levels of endangerment for each of the test file formats. Additionally, I asked them to write a brief justification for each of their choices. I compiled participants answers into one document and shared this document with participants. I asked participants to review the document and then asked them to re-rate the file formats after considering the judgments and justifications of their fellow expert participants.

I addressed the second step of my outlined methodology by similarly employing the Delphi method. I presented a second group of experts with a list of file format evaluation factors compiled from a dozen file format evaluation factor lists found in the literature, and I asked them to rate the factors according to their relevance as a cause of file format endangerment. I asked them to explain the rationale for their answers as well as their thoughts on how the factor should be measured and how data should be collected for the factor. As with the file format endangerment rating exercise, I compiled participants answers into one document, shared this document with participants, and then I asked them to re-rate the factors after considering the judgments and justifications of their fello participants. During the first Delphi round, I asked participants to suggest additional factors that were not on the list that they believe should be included, and why. During the second round, I presented them with any non-redundant factors suggested in the first round as well as any factors that may emerge through analysis of the justification data collected during the first Delphi study.

Lastly, I brought the elements of the two Delphi studies together in a final phase of data collection. During this phase, a trained reviewer used the top-rated factors from the second phase and applied them to the list of file formats from the first phase. Making use of a guide comprised of information collected from Delphi participant comments, the reviewer then rated each file format's level of endangerment based on the collected factor data, and then rated each factor for relevancy as a cause of file format endangerment. I interviewed the reviewer at the end of this process to collect feedback on the process he used to collect information for each factor, how useful they found each factor to be in assessing file format endangerment, and any other thoughts and opinions they had about the process.

1.2. Expected Contributions

Through answering these research questions I will contribute to digital preservation research in several ways. First, I have collected expert opinions on the endangerment levels of the fifty test file formats that I can share with the community. The digital preservation community may use this information to guide their digital preservation decisions. Second, I have collected information that informs the selection of indicators for a file format endangerment index that can provide much needed file format risk assessment information to the digital preservation community. Finally, I have performed initial tests of the resulting proposed file format endangerment index factors and have built a foundation from which future file format endangerment research may develop.

CHAPTER 2: LITERATURE REVIEW

This literature review examines the key areas of research and development that inform the design of the study conducted and discussed here. First, I cover the key conceptual and theoretical foundations that have shaped digital preservation and file format research and practice over the years. Then I cover the discussion in the literature about the challenge that file formats pose to digital preservation. I then examine existing research and development of file format management tools and demonstrate the gap in file format risk analysis within the digital preservation research and development landscape. Next, I review the criteria for evaluating file formats previously published in the literature. Then, I explore how methods from conservation biology can be applied to the problem of file format endangerment analysis. Finally, I review index development methods and how they will be applied in this study.

2.1. Conceptual and Theoretical Foundations

Over the past several decades, researchers and practitioners have developed a number of theoretical models and concepts that have had a clear impact on digital preservation research, practice and tool development. Three such contributions are the Reference Model for an Open Archival Information Systems (OAIS), the concept of Significant Properties, and the Digital Curation Lifecycle.

The Consultative Committee for Space Data Systems (CCSDS) released the Reference

Model for an Open Archival Information Systems (OAIS) as a "technical Recommendation for use in developing a broader consensus on what is required for an archive to provide permanent, or indefinite long-term, preservation of digital information" (2002, p. iii; 2012, p. iii). The OAIS reference model tied together critically important concepts that have informed digital preservation research and practice throughout the years, and as Lee stated, "the OAIS has come to be a widely assumed basis for research and development on digital archiving" (2005, p. 4). The elements of the reference model that are most relevant to this research are the *Information Object*, the *Representation Information Object*, and *Transformational Information Properties*, which best illustrate the function of the file format in digital preservation.

According to the OAIS, an information object is, "composed of a Data Object that is either physical or digital, and the Representation Information that allows for the full interpretation of the data into meaningful information" (CCSDS, 2012, p. 4-20–4-21). These relationships are illustrated in *Figure 2.1.1. Representation Information* is the information that accompanies the digital or physical object and is necessary to interpret or render the content of the object.

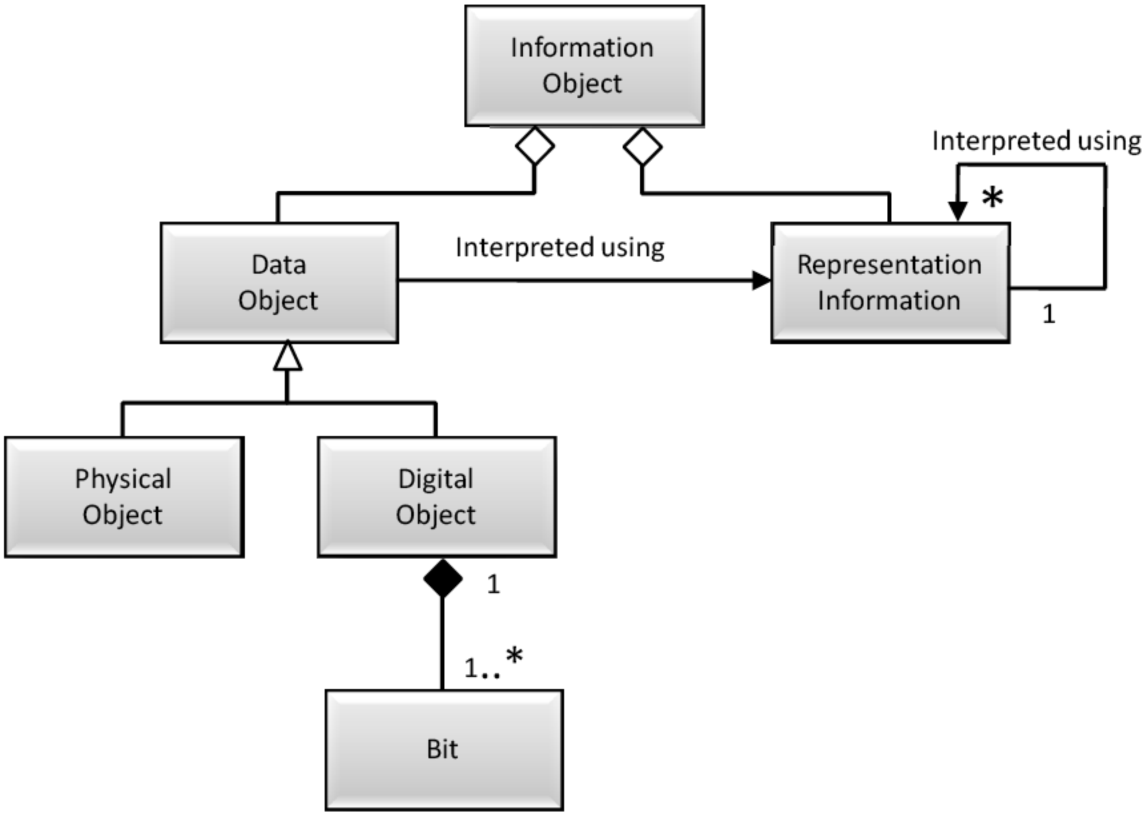


Figure 2.1.1 OAIS Information Object. (CCSDS, 2012, p. 4-21).

There are three different types of Representation Information: *Structure Information*, *Semantic Information*, and *Other Representation Information*. *Structure Information* provides the structural information necessary to translate the bit sequences of the digital object into a form that is meaningful and understandable to humans. *Semantic Information* provides additional meaning to the structural information such as the language used in the information object. *Figure 2.1.2* illustrates these relationships and also illustrates the possibility that *Other Representation Information* can be part of the *Information Object*. According to the OAIS, “information defining how the Structure and the Semantic Information relate to each other, or software needed to process a database file would be regarded as Other Representation Information” (2012, p. 4-22).

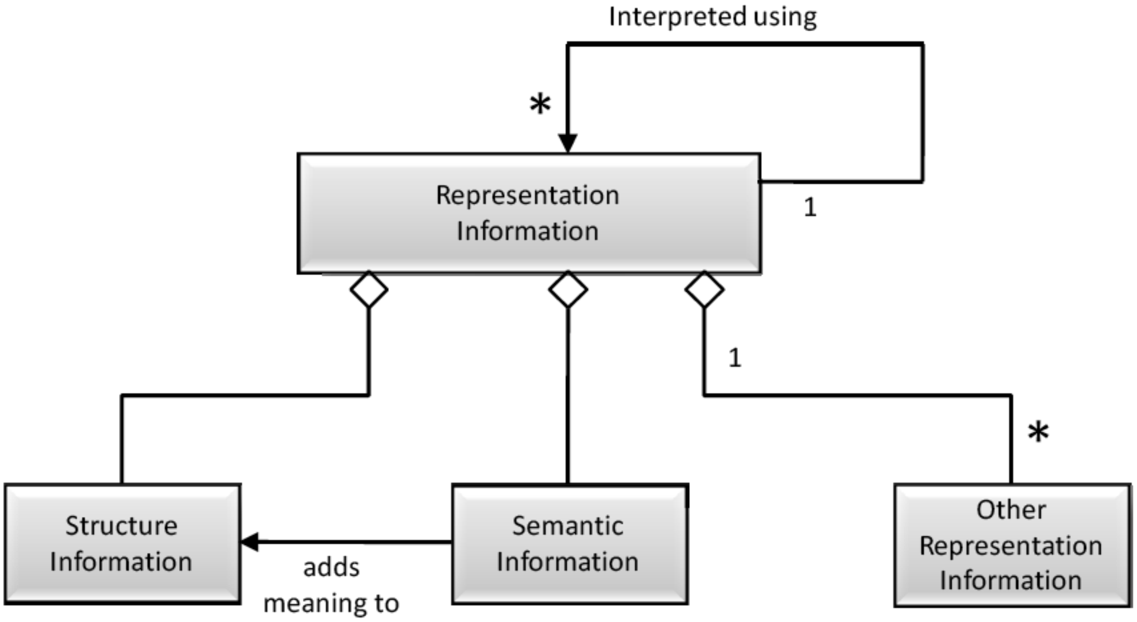


Figure 2.1.2. OAIS Representation Information Object. (CCSDS, 2012, p. 4-23).

A notion introduced in the 2012 version of the OAIS Reference Model is *Transformational Information Properties*, or what are often referred to as *Significant Properties*. Hedstrom, Lee, Olson, and Lampe refer to significant properties as the, "features, attributes, or properties that impinge upon future use and understanding" (2006, p. 161). Significant properties can be things like the "look and feel," functionality, and behavior of the digital content.

Reagan Moore wrote that digital preservation is "communication with the future" and a major challenge of preservation is to "incorporate new technology effectively, while conserving preservation properties such as authenticity, integrity, and chain of custody"

(2008, p. 64). Allison, Currall, Moss and Stuart wrote that, "Even if the bitstream has been dignified with an ISBN or ISSN number, it does not exist like a printed book in multiple identical copies but only in multiple almost certainly nonidentical renditions. Across the board, the bitstream is the constant, although behaviors are not" (p. 368).

Significant properties are important when considering file format endangerment levels. While there may be software available that can render the information stored in a particular file format, that software might not faithfully reproduce all of the significant properties that maintain the information's authenticity and integrity. In certain contexts significant properties can be the line between whether or not meaningful access to a digital object has been preserved. This distinction has a definite impact on a file format's endangerment level.

David Levy describes long-term digital preservation as a "socio-technical accomplishment" where a balance must be struck between the technological demands and the need to focus on the human and organizational factors that affect digital preservation (1998). The two are inextricably linked in that the technological aspects of preserving access to digitally encoded information are created by and must be managed by people. It is people who are charged with making the best digital preservation decisions so that not only are the bitstreams preserved, but so too are the significant properties, authenticity, and other contextual information that maintain a digital object's value. In the context of file formats, decisions must be made around perceived and actual risks associated with particular file formats, but it is always best to make decisions on the actual risks.

2.2. The Problem of File Formats in Digital Preservation

In terms of the OAIS Reference Model, file formats are a part of the structural component of representation information, and maintaining access to digitally encoded information is dependent on the presence of representation information. Preserving access to digitally encoded information is a challenge due to the continually changing nature of the technology that people use to produce and render it. The literature portrays an evolving discussion of the problem of file formats in digital preservation over four decades.

In 1971, Dollar cited having documentation on the software used to create and manipulate machine-readable records as a necessity to continued digital data access:

> …A computer program directory would at the least describe the programs used to manipulate the data and the computer employed. For the twenty-second and twenty-third centuries a source listing of the programs and a flow chart should be included (p. 30).

Margaret Hedstrom also discussed the challenges of archiving machine-readable records due to “software dependence”:

> The need for special software to access a data file and retrieve information from it is a major obstacle to transfer, processing, and distribution of a data set. Some software-dependent data sets are unusable if transferred to the archives without companion software (1984, p. 44).

In 1992, Michael Lesk wrote at length of the problems created by file formats and the constant change in the software that creates them. He pointed out, “Format, software and hardware are often intermingled: information may be preserved but if the software to print, search, and edit it has gone, it may be quite costly to make any use of it” (New Media are a Problem, ¶6). Note that he did not say that the bitstreams cannot be preserved, just that it could be expensive to find a way to render them. He went on to speak more specifically about the challenge of file formats: “Unfortunately there is also a much wider variety of logical formats, and much more varied expertise is required to deal with the software content than with the physical material” (New Media are a Problem, ¶6).

In 1994, David Bearman made note of the same problem when he wrote, “Electronic records are always virtual documents, that is they exist under software control and are dependent on some hardware, even if they are (someday) truly ‘inter-operable’ across hardware platforms. Because a generation of hardware and software (the length of time before obsolescence) is less than five years and because storage media generations are equally volatile, the electronic records must be regularly migrated to new hardware, software, and media” (1994a, p.21).

Soon after, Paul Conway wrote on the obsolescence of data retrieval systems. He said, “Digital storage media must be handled with care, but they most likely will far outlast the capability of systems to retrieve and interpret the data stored on them. Since we can never know for certain when a system has become obsolete, libraries must be prepared to migrate valuable image data, indexes, and software to future generations of the technology” (1996, Priorities for Action, ¶1). Here he spoke of never knowing when a system becomes obsolete, and he suggested that libraries must be ready to take action to preserve access to their digital collections. It is important to note this sense of uncertainty around when and if the risk will materialize.

Margaret Hedstrom rang an alarm bell when she said, “More insidious and challenging than media deterioration is the problem of obsolescence in retrieval and playback technologies. Innovation in the computer hardware, storage, and software industries continues at a rapid pace, usually yielding greater storage and processing capacities at lower costs" (1998, p. 191). In the same publication Hedstrom suggested that, among other things, a digital object's file format should serve as a basis for what an institution can invest in digital preservation initiatives. The implication here is that if an institution cannot guarantee access to the information encoded in a particular file format, it may not, or even should not spend time and money on trying to preserve digital objects stored in that format.

In reference to the longevity of file formats, Jeffrey Rothenberg famously said, "digital information lasts forever—or five years, whichever comes first" (1999b, p. 2). In another publication from the same year, he wrote, "Digital documents are vulnerable to loss via the decay and obsolescence of the media on which they are stored, and they become inaccessible and unreadable when the software needed to interpret them, or the hardware on which that software runs, becomes obsolete and is lost" (1999a, p. v).

Software and file formats continued to be cited as core sources of digital preservation trouble through the 2000's. In 2000, Seamus Ross said, "Access to material created using superseded operating systems (e.g. CP/M) or word-processing (e.g. WordStar) and database (e.g. Dbase III) applications is difficult…resources created in digital form are fragile and easily prone to becoming physically and logically inaccessible" (p. 12). In 2002, Kenneth Thibodeau named file formats as the "starting point for all digital preservation."

A sense of urgency around file formats is instilled in the statement:

> Because of the peculiarities of the problem, however, we cannot afford to wait any longer: New digital formats are being invented all the time. Most of them, like their hard- and software environments, live a short life until they are supersede by new formats. Access to documents stored in an obsolete format is restricted at best" (Borghoff, Rödig, Scheffczyk, & Schmitz, 2006).

In 2010, file formats and the software that create and read them were again highlighted as one of the premier challenges in digital preservation. The National Digital Information Infrastructure and Preservation Program (NIIIPP) reported, “Content formats can be complex and fragile. They are often not well documented and frequently become obsolete” (p. 12). Kurt Bollacker agreed that, “with all digital media, a machine and software are required to read and translate the data into a human-observable and comprehensible form. If the machine or software is lost, the data are likely to be unavailable or, effectively, lost as well” (2010, p. 106).

In contrast to this, Sarah Higgins had a more positive perspective of progress in the field: “After a period of definition and consolidation, the subject now boasts a growing international professional base, a developing research agenda, practical tools and collaborative projects and a workforce trained to Higher Education level” (2011, p.11). However, the majority of the literature reflects a darker view. This includes Seamus Ross, as reflected in several comments he wrote in 2007:

> The preservation community has not yet carried out sufficient underlying experimental and practical research either to deliver the range of preservation methods and tools necessary to support preservation activities or to provide us with sufficient data to reason effectively about preservation risks or how to manage them (p. 7).

An early but influential report in 1996 conveyed that technological aspects of digital preservation still needed to be researched and addressed. “Even after more than forty years of growth, the digital world of information technology and communication is still relatively young and immature in relation to the larger information universe, parts of which have been under development for centuries” (Task Force on Archiving of Digital Information, 1996, Need for Deep Infrastructure, ¶1). Without focused research in the finer points of digital preservation science, we will not have a strong enough collection of solutions.

Little data has been collected about the risks file formats may pose, and consequently, it is challenging to develop solutions-based research. It is necessary to implement continued data collection so that we may better understand the problem and to better understand where the risks lie. Seamus Ross summed up this sentiment:

> Not only do we need to try to better understand what we might do to alleviate obstacles to the longevity of digital materials, we must do more to define the uncertainties related to digital preservation and to convert these uncertainties into known, measurable and mitigatable risks. We should, of course, make a genuine distinction here between perceived risk and 'actual' risk; an actual risk represents an assessed and measurable risk – we just do not know in a measurable way in the context of digital objects which risks are actual (2007, p.17).

A report from DigitalPreservationEurope from the same year stated:

> Appropriate metrics for various types of risks as well as for their economic and other consequences have to be defined. Algorithms should also be developed that support the measurement of various types of risks on the basis of such metrics. … Furthermore, methods of knowledge representation and reasoning can be utilised to represent these risks in an explicit, machine readable form that can be automatically processed and analysed by applying methods of machine learning and automated reasoning (2007, Risk, ¶1).

Not everyone believes that file formats and the technological advances in rendering software pose a risk to our ability to access digital information over time. David Rosenthal and Chris Rusbridge both published papers that brought into question the degree of severity the community believed the problem possessed.

Rosenthal wrote, "format obsolescence is not a significant threat to the overwhelming majority of digital content we wish to preserve" (2010). He went on to say, "If we ask 'what would have to happen for these formats no longer to be renderable?' We are forced to invent implausible scenarios in which not just all the independent repositories holding the source code of the independent implementations of one layer of the stack were lost, but also all the backup copies of the source code at the various developers of all these projects, and also all the much larger number of copies of the binaries of this layer" (p. 206). He goes on to argue that file format obsolescence is technically impossible for most file formats:

> Any format that has an open-source renderer is effectively immune from format obsolescence because there is no plausible scenario in which it will stop working in the current environment and, if it does, the environment in which it did work can be re-created. Further, any format that can be rendered by a binary plugin for an open source environment can be made immune simply by preserving the bits of the binary plugin…Thus the practical questions about the obsolescence of the formats used by today's readers are really how convenient it will be for the eventual reader to access the content, and how much will be spent when in order to reach that level of convenience (p. 207).

Ultimately, he says, "format obsolescence is a rare problem that happens infrequently to a minority of unpopular formats" (p. 208). Chris Rusbridge agreed that file format endangerment does not pose the immediate threat that the community thinks. He wrote, "File formats become obsolete rather more slowly than we thought" (2006).

The literature demonstrates the importance of file formats to a large portion of digital preservation community. It also demonstrates the need to clarify the nature and degree of any risks that may be associated with accessing information encoded in particular file formats over time.

2.3. File Formats in Digital Preservation Research and Development

Research and Development in file formats in digital preservation has focused on the development of file format identification tools, file format registries, risk notification tools, and systems that integrate two or more types of file format tools. My examination of the digital preservation software development landscape reveals progress in file format identification and digital preservation system development, but little forward movement in developing processes for systematic file format risk assessment.

2.3.1 File Format Identification Tools

The Harvard University Libraries (HUL) explained the value of knowing the format of a digital object:

> The format of a digital object must be known in order to interpret the information content of that object properly. Without knowledge of its format, a digital object is merely a collection of undifferentiated bits. Thus, format typing is fundamental to the effective use, interchange, and preservation of all digitally-encoded content (2008, p. 1).

In some cases it is difficult to accurately identify the format of a digital object, particularly in the context of large, heterogeneous digital collections. As a result, a number of developers created tools to aid in identifying file formats. Some of these tools include the UNIX **file** utility, Digital Record Object Identifier (DROID), JSTOR/Harvard Object Validation Environment (JHOVE), Format Identification for Digital Objects (FIDO), File Information Toolset (FITS), TrID, Filereg, Mfile, Validate, File Fingerprints, Token Intersection, and Probabilistic Token Validation.

The UNIX **file** command is the predecessor to all of the file format identification tools discussed here. The **file** command has been used to identify file formats since 1973 (Underwood, 2009). The file format identification tool, DROID was created by The National Archives (TNA), of the UK in order to "extend the functions of the PRONOM technical registry [discussed below] by performing automated batch identification of file formats" (The National Archives [TNA], n.d.-a). JHOVE is a file format identification and characterization tool developed in a cooperative effort between Harvard University Libraries (HUL) and Journal Storage (JSTOR). Marco Pontello created a lesser-known file format identifier called, TrID (Pontello, n.d.).

Format Identification for Digital Objects (FIDO) (Open Planets Foundation, 2012) is a relatively new tool, first released in December 2010 with an additional version released in March 2011. The File Information Toolset (FITS) is a wrapper around several standalone file identification and characterization tools: JHOVE1, Exiftool, National Library of New Zealand Metadata Extractor, DROID 3.0, FFident, and the UNIX File command (McEwen & Goethals, 2009). As part of the File Format Identification research project sponsored by the MITRE Corporation (2009), the MITRE team developed several prototypical file format identification tools to aid in digital file forensics. These tools are *Filereg*, *Mfile*, *Validate*, *File Fingerprints*, *Token Intersection*, and *Probabilistic Token Validation*.

2.3.2 File Format Registries

File format registries are designed to collect and share information about file formats, the software that can be used to render them, the risks associated with particular file formats and what actions can be taken if the formats are at risk of being unrenderable. Notable file format registries are PRONOM, the Global Digital Format Registry (GDFR), the Unified Digital Format Registry (UDFR), and the Archive Team's "Let's solve the file format problem" wiki.

PRONOM is a "database of the technical components necessary for accessing and processing electronic records" (Brown, 2005b). It contains information about file formats, software products, operating systems, hardware components, and storage media. It was first released in 2004 by The National Archives (TNA) of the United Kingdom as a component of its government dataset preservation service (Brown, 2007). In each file format entry, there is a field designated for "format risk," but inspection of the live database reveals that there is currently no risk information available for the file formats (TNA, n.d.-b).

The Global Digital Format Registry (GDFR) is a similar project that was initiated in 2002 as a result of discussions between team members of the Harvard Library Digital Initiative and MIT DSpace project (Abrams & Seaman, 2003). Like PRONOM, the GDFR aims to collect and share representation information about digital file formats, but what distinguishes it from PRONOM is that it aims to "define a common network protocol by which multiple independent, but cooperating, registries can communicate with each other and synchronize their holdings of format representation information" (Abrams & Flecker, 2005). In other words, instead of being a standalone database it is intended to ingest and manage format data from a distributed network of databases, of which PRONOM data could be one of many components.

The GDFR team created a data model, originally informed by the PRONOM 4 information model (Harvard University Libraries, 2006), but adapted to include more granular information on the file formats that can be used to monitor and address file format endangerment issues. For example, the data model includes space for human or corporate agent information associated with a file format. This information could be used to contact parties with the knowledge and ability to provide access to information stored in a particular file format if rendering software has become unavailable.

In 2009, the Unified Digital Formats Registry (UDFR) Working Group announced their intentions to form a Unified Digital Formats Registry that combined the contents of TNA's PRONOM and the GDFR into one central repository (Unified Digital Formats Registry [UDFR] Working Group, 2009b). This new registry is technically based on PRONOM's existing architecture and database, but has been expanded to include the data in the GDFR registry. The model for the registry is "based on shared governance, cooperative data contribution, and distributed data hosting" (Unified Digital Formats Registry Working Group, 2009a). The project was completed in the summer of 2012. Unlike PRONOM and the GDFR, UDFR allows contributions from the public. It has an open contributor policy that states:

> There are no prescriptive requirements for contributor eligibility other than providing minimum personal information: name, email address, and institutional affiliation and job title. Instead, the UDFR relies on strong provenance and complete change history at the level of each individual assertion (Regents of the University of California, 2012).

Challenges common to many of the funded format registry projects, particularly GDFR and UDFR, include the problem of ongoing maintenance and development. In a 2013 publication, McGath reported no ongoing activity for either project. In contrast to this more traditional model, the "Let's solve the file format problem" project seeks to provide "an institution-neutral, public-domain, easy to navigate site containing this information, the "problem" can be addressed both by users of the Wiki and the many, many related attempts to achieve this goal" (Archive Team, 2012). Continuing activity on the Archive Team wiki suggests but does not prove that this alternate model may be a more sustainable means of collecting and providing access to file format information.

2.3.3 Risk Assessment and Notification Tools

Several projects have approached the process of file format risk assessment and notification. These are the Automated Obsolescence Notification System (AONS), AONS II, parts of the Archive Ingest and Handling Test (AIHT), Plato, Scout, and research conducted at the Austrian Institute of Technology.

AONS was a project of the National Library of Australia (NLA) and the Australian

Partnership for Sustainable Repositories (APSR) and built upon work of the Preservation Architecture for New Media and Interactive Collections (PANIC) project, discussed later. In 2006, AONS was developed to create a file format obsolescence alert system, specifically for the DSpace digital repository platform. The alert system was to be built on an architecture that used DROID for file format identification, and PRONOM and Library of Congress Directory of Formats to provide obsolescence risk evaluation. If file formats found in the repository are identified to be at risk, the system generates a risk report and sends the report to the repository manager (Australian Partnership for Sustainable Repositories, 2006).

In 2007, work on AONS II began in order to refine the AONS services. Notably, the AONS II report stated, "an initial business driver for the project was a perceived need for a tool which could automate much of the assessment process, using standardized metrics that would support machine-formulation of recommendations on risk levels" (Pearson & Webb, 2008). Unfortunately, the project relied heavily on risk reporting capabilities of PRONOM, which have yet to come to fruition. Since the AONS II report was issued, there has been no further development of AONS.

The Archival Ingest and Handling Test (AIHT) project was funded by the Library of Congress to "assess the digital preservation infrastructures of four small, real-world digital archives" (Anderson, Frost, Hoebelheinrich, & Johnson, 2005, ¶1). The four partners were Johns Hopkins University, Sheridan Library; Harvard University Library; Old Dominion University Department of Computer Science; and Stanford University, Libraries and Academic Information Resources (Library of Congress, n.d.). As part of the AIHT, the Stanford University participants developed a file format risk-assessment system. They based their system on JHOVE for file format identification and representation information and the

Arms and Fleischhauer (2005) list of preferred file formats, from which they created a matrix for risk-assessment. From this they developed what they call the Empirical Walker Process, intended to be a fully automated metadata and risk-assessment generator that flags materials that may be in danger of becoming obsolete (Anderson, Frost, Hoebelheinrich, & Johnson, 2005).

After developing this prototype system, Anderson, Frost, Hoebelheinrich, and Johnson evaluated the resources required to automate and maintain a preservation assessment of the Empirical Walker Process, such as maintaining the infrastructure to support the process. While they have yet to fully develop this process, they suggested that the cost to manage such a system was too much for one institution to bear and suggested that "perhaps a federated approach to some of this activity, as a service to a community of repositories and their users, would be most economical" (2005, General Conclusions, ¶2).

Plato was developed as part of the Planets preservation-planning project. Plato addresses many aspects of preservation planning (Becker & Rauber, 2011). Among them is assessing file format criteria that could indicate risk. They propose to evaluate file formats based on the criteria: browser support, standardization, ubiquity, stability, licensing, compression, format documentation, tool support, comparative file size, complexity, disclosure, master can be used as access copy, Optical Character Recognition (OCR applicable, and adoption. Becker and Rauber cite several obstacles toward realizing the goal of automating the process of measuring and evaluating formats based on these criteria: 1. only roughly 20% of the criteria can be automatically measured, 2. external sources of data or not complete and, 3. there is a lack of standardized benchmarks that can be used in comparative analysis.

Scout is a semi-automatic preservation watch system being developed within the Scalable Preservation Environments (SCAPE) project, "an EU-funded project which is directed towards long term digital preservation of large-scale and heterogeneous collections of digital-objects" (SCAPE, n.a.). Scout was designed to collect information from various sources that can be used to detect risks to digital content. It collects information from various registries like PRONOM as well as through natural language extraction from the World Wide Web (Faria, 2013; Faria, et al, 2013). This tool is still under development and has undergone only basic, proof-of-concept testing.

Another, similar approach toward file format risk analysis is being developed by Roman Graf and Sergiu Gordea (2013), both of the Austrian Institute of Technology. They are also developing a system that collects data from various sources to analyze file formats for what they call, "preservation friendliness." They designed their system to collect data from PRONOM, DBPedia, and Freebase on twenty-one identified risk factors:

1. Software Count
2. Vendors Count
3. Versions Count
4. Has Descriptions
5. Has MIME type
6. Existence Period
7. Is Complex Format
8. Is Wide Disseminated
9. Is Outdated or Deprecated
10. Has Genre
11. Has Homepage
12. Is Open (Standardised)
13. Has Creation Date
14. Has File Migration Support

15. Digital Rights Information
16. Has Publisher Information
17. Has Creator Information
18. Is Popular Format
19. Has Compression Support
20. Supported by Web Browser
21. Has Vendor Support

They collected and analyzed data for these factors for a set of thirteen representative file formats to produce a total risk percentage value for each file format.

2.3.4 File Format Risk in Digital Preservation Systems

A few groups have developed digital preservation systems that incorporate file format risk analysis into workflows. These are the Preservation Services Architecture for New media and Interactive Collections (PANIC), Ex Libris' Rosetta, Tessella's Safety Deposit Box, and the National Library of the Netherland's (KB) *e*-Depot.

PANIC is a "semi-automated digital preservation system based on semantic web services" (Hunter & Choudhury, 2006, p. 1). The project, funded by the Cooperative Research Centre for Enterprise Distributed Systems Technology (DSTC) and the Australian Federal Government's CRC Programme, facilitated the building of a prototype system to assess a digital object's obsolescence risk and subsequently invoke migration or emulation tools to counteract the risk. The system architecture contains invocation, notification, discover, and provider components. The invocation component was designed to detect obsolescence using information retrieved from the built-in software version registry via a notification agent. This registry contains information about software that is used to render the objects in the collection. Once notified of risk, the discovery component is set into action to locate appropriate preservation services using the OWL-S ontology that is used for describing and discovering web services. The provider component then sends the at-risk files to the located service which then performs the requested service (Hunter & Choudhury, 2004). There has been no development of PANIC beyond the prototype phase.

Rosetta is a digital preservation system produced by the Ex Libris Group (2010). The system has a deposit module, a working area, a permanent repository module, an operational repository, a preservation planning module, an administration module, and an access module. According to the software description, the preservation-planning module provides risk analysis of file formats, but there is no indication as to how this is accomplished. I contacted a representative of Ex Libris who stated that due to the proprietary nature of their product, they could not share information beyond what is available online.

Safety Deposit Box (SDB) is a digital preservation system developed by Tessella (2013). Key features of SDB are ingest, data management, storage, access, preservation planning and action, and administration. The preservation planning and action feature uses file characterization tools to assess file format risk, though there is no clear source of internal or external file format risk information and no clear evidence that this function is operational. As of this writing, the file format evaluation component of SDB is still not production ready, though, "Tessella are moving to a 'linked data' registry in the next release. The plan is to revisit the ability to define a format risk assessment in a future release once the linked data version is stable" (Evans, M., personal communication, January 24, 2014).

e-Depot is a system built for the National Library of the Netherlands using the IBM system, Digital Information Archiving System (DIAS) (Oltmans, van Diessen, and van

Wijngaarden, 2004). DIAS was extended to include a Preservation Subsystem that included a functionality called the Preservation Manager that stores technical metadata that specifies the software and hardware necessary to render the file formats stored in *e*-Depot. This functionality was designed to meet three objectives: "1) Identify[ing] the electronic publications in danger of becoming inaccessible due to technology changes, 2) Planning the activities associated with preservation, i.e. implementing migration and/or emulation strategies, and 3) Specifying the software and hardware environments required to render an electronic publication" (p. 281). At the time of this writing, the KB web page on *e*Depot states that, "Preservation functionality will be enhanced in future DIAS versions to generate signals when stored assets must be converted or migrated to ensure their availability" (Koninklijke Bibliotheek, n.d.). Attempts to communicate with representatives from the KB to learn more yielded no results.

Digital preservation researchers and developers have put a great deal of work into creating tools and systems that are designed to manage and preserve digitally encoded information. A close examination of the existing tools, however, reveals a gap in a critical area of need: none of these tools and systems actually addresses the issue of file format risk monitoring, though some developers claim their systems do or will in the future. Many of the tools and systems discussed here claim that their file format risk analysis components will come from PRONOM, but PRONOM does not currently contain information on file format risk information. In fact, none of the tools or systems listed here has proven functionality in file format risk analysis. This shows that though the digital preservation community indicates that it is important to monitor file format risk, they have yet to find a viable way to do this.

2.4. File Format Assessment Criteria

Effective analysis of file format endangerment requires a well-constructed and validated index to guide data collection. The key to creating a valid index is choosing the right factors that cause the phenomenon measured. Previously, researchers from various institutions created several different lists of file format evaluation criteria. Some of these lists of criteria were designed to evaluate aspects of file formats that can contribute to or alleviate risks that can prevent access to information encoded in a particular file format. While none of these lists were created with the intention of creating a file format endangerment index, the approaches used are similar enough to provide a useful starting point for the index development process.

I have identified twelve sets of file format evaluation criteria from the literature. *Table 2.4.1* lists the criteria; the projects, programs, and institutions with which they are associated; and citations to the literature in which the criteria are discussed. These lists of evaluation criteria were the basis from which I began this study.

Project/Program/Institution	File Format Criteria	Source
Automated Preservation Assessment of Heterogeneous Digital Collections (AIHT)	• Adoption • Disclosure • Transparency • Self-Documentation • External Dependencies	Anderson, et al, 2005; NDIIPP, 2005
Internetbevaringsprojektet (the Internet Preservation Project); Statsbiblioteket (The State Library), Det Kongelige Bibliotek (Royal Library, Denmark)	*Openness* • Open, publicly available specification • Specification in public domain • Viewer with freely available source	Kongelige Bibliotek, 2004a; 2004b

Project/Program/Institution	File Format Criteria	Source
	• Viewer with General Public Licensed (GPL) source • Not encrypted *Portability* • Independent of hardware • Independent of operating system • Independent of other software • Independent of particular institutions, groups or events • Widespread current use • Little built-in functionality • Single version or well-defined versions *Quality* • Low space cost • Highly encompassing • Robust • Simplicity • Highly tested • Loss-free • Supports metadata	
INvestigation of Formats based on Risk Management (INFORM)	• Whether or not royalties or license fees are or may be requested • Whether the source or specification can be independently inspected • Whether revisions have maintained support for backward compatibility • Whether it is complex or poorly documented • Whether it is widely	Stanescu, 2004

Project/Program/Institution	File Format Criteria	Source
	accepted or simply a niche format • Whether competing or similar formats or components exist • Whether embedded metadata can be mapped to other formats • Whether digital rights management (DRM) encryption or digital signatures can be used • Whether applicable expertise can be easily found • Whether revisions happen so fast that the archive cannot keep up with demand • Whether extensions, such as executable sections or narrowly supported features, can be added • Whether authenticity can be easily compromised during transformations, either accidentally or maliciously • Whether the associated organization or community is too small, in danger of collapsing, unique in its class or not easily replaceable	
International Research on Permanent Authentic Records in Electronic Systems 2 (InterPARES)	• Widespread use • Non-proprietary origin • Availability of specifications	InterPARES, 2007

Project/Program/Institution	File Format Criteria	Source
	• Platform independence (interoperability) • Compression	
Koninklijke Bibliotheek (KB)	• Openness • Adoption • Complexity • Technical Protection Mechanism (DRM) • Self-documentation • Robustness • Dependencies	Koninklijke Bibliotheek, 2008
Math*Diss* International Project and EMANI project; Niedersächsische Staats- und Universitätsbibliothek, Götingen	• Error tolerance • Long term stability • Full open specification • System dependence • Ease of handling • Independence of commercial interests and influence	Fischer, 2003
The National Archives (TNA-UK)	*Popularity* • Ubiquity • Numbers of viewers available • Disclosure • Stability and backwards compatibility *Utility* • Loss of "functionality" caused by migration • Retention of metadata on migration • Proportion of digital objects/records • Support for redaction *Technical Features* • Ease of identification • Technical dependencies	The National Archives, 2005a, 2005b

Project/Program/Institution	File Format Criteria	Source
	• Transparency • Probability of information loss due to corruption • Likelihood of infection • Ease of validation • Availability of migration path • Losslessness *Commercial Factors* • Cost of migration away from the format • IPR impacts • Verbosity	
National Centre for Radio Astrophysics	• Developer of file format goes out of business • Developer stops supporting that format • The market share of the developer declines • Supporting program of the software change significantly • Third party support is lacking • Format depends on obsolete hardware or operating system • Format is proprietary • New versions of application software may not support earlier format versions • Application software developers do not release documentation or release accurate documentation • Software too complex and document structure layout is	Barve, 2007

Project/Program/Institution	File Format Criteria	Source
	proprietary	
Groupe Pérennisation des Informations Numériques (PIN)	• Capability of the format to represent all the information whose long term preservation is to be ensured • Publicly standardized format • Possibility of modifying the data at a later date • Tools and creation and manipulation facilities • Complexity /Simplicity of the format • Inspectability • Metadata • Performance of available implementations	Huc et al, 2004
Preservation and Long-term Access through Networked Services (PLANETS)	• Ubiquity • Support • Disclosure • Document Quality • Stability • Ease of Identification • Ease of Validation • Use of Compression • Intellectual Property Rights (IPR) • Complexity	Becker, et al, 2008
Risk Management for Digital Information Project; Council on Library and Information Resources	• Content fixity • Security • References • Cost • Staffing • Functionality • Legal	Lawrence, et al, 2000; Rieger and Kenney, 2000
Service Oriented Architecture (SOA); University of Minho, Portugal	• Market share • Support level • Is standard	Ferreira, 2006, 2007

Project/Program/Institution	File Format Criteria	Source
	• Open specification • Supports compression • Lossy compression only • Supports transparency • Embedded metadata • Royalty-free • Open source • Backwardly compatible • Documentation level • Competing formats available • Digital Rights Management (DRM) support • Update frequency • Supports custom extensions • Life time • Transparent encoding • Reader single producer • Single reader • Open source reader • Multiplatform reader	

Table 2.4.1. File format evaluation criteria from the literature.

The AIHT was a project funded by the National Digital Information Infrastructure and Preservation Program (NDIIPP) at the Library of Congress to generate knowledge of digital preservation processes through the practical applications at four participating institutions: Harvard, Johns Hopkins, Old Dominion, and Stanford Universities. As part of the Archive Ingest and Handling Test (AIHT), members of the Stanford University team adapted file format evaluation criteria created by Carl Fleischhauer and Carolyn Arms into their own matrix for an automated file format risk assessment methodology (Anderson et al, 2005;

Shirky, 2005). In this matrix, they used the factors of *disclosure*, *adoption*, *transparency*, *self-documentation*, and *external dependencies* to analyze the risk of several file formats. They chose not to include the factors, *impact of patents* and *technology protection measures*. They cite the complexity of patents' impact on file formats as a reason not to include the factor in their matrix, and decided that technology protection mechanisms can affect individual files of any file format and not a particular file format as a whole, so this factor was also not included.

According to Arms and Fleischhauer, *disclosure* refers to the degree to which specifications are available for use in maintaining access to the digital content of the file format. The factor, *adoption*, refers to how widely used the file format is. The *transparency* factor is related to how accessible the file format is for analysis and rendering using basic tools like text editing software. *Self-documentation* refers to the extent to which descriptive, technical, and administrative metadata can be included in the file format. Lastly, the *external dependencies* factor indicates the degree to which the file format depends on external relationships with things like hardware, operating systems, or software.

Det Kongelige Bibliotek (The National Library of Denmark) issued a report on their review and handling of file formats. As part of this review, they reviewed several file format evaluation criteria (2004). The individual criteria are grouped under the categories of *openness*, *portability*, and *quality*.

The criteria under the *openness* category refer to whether the file format has open and publicly available specifications that can be used for rebuilding or reverse engineering rendering software. It also examines whether or not this specification is in the public domain and that the specification is not bound by patents or copyright issues that prevent future use.

The *openness* criteria address whether or not viewer software is currently available and whether it is hampered by licenses that could prevent its use in the future. Lastly, the *openness* criteria address whether or not the format requires a special encryption key to be read.

The *portability* criteria suggest that a file format should not be dependent on particular hardware, operating systems, or other software. Additionally, the file format should not be dependent on particular institutions, groups, or events; a file format that has been created for a particular institution, group, or event may have peculiarities that make it more difficult to render in the future. The *portability* criteria also state that a format should be in widespread use, should have little built-in functionality, and should have either a single version or well-defined versions.

The criteria included under the *quality* category suggests that a high quality file format will not take up much storage space, is highly encompassing ("can be used as a target for a greater number of other formats," i.e., can be rendered), robust (is not easily corruptible if bits in the format "flip"), is simple, highly tested, is loss-free (does not lose substantial amounts of information if other formats are converted to it), and supports metadata (allows for the storage of metadata within the file).

As the author suggests in this report, many of these criteria are subjective and many of them conflict with one another. They were designed to be issues to consider while determining whether or not to use or convert a file format and not as an all-or-nothing checklist. Nonetheless, there are 19 criteria listed under the three categories; some of which are straightforward and easy to assess, and some of which are vague and would be difficult to apply. Assessing whether or not a file format is what they call, "robust" will be easier to determine than if it is “highly encompassing.”

In a report on assessing file formats for the INvestigation of Formats based on Risk Management (INFORM) methodology project, Andreas Stanescue listed 13 risks to the durability of a file format (2004). The INFORM methodology was developed as a product of the Online Computer Library Center (OCLC) to measure risks to file formats and to provide guidance for risk mitigation plans. Within the INFORM methodology, OCLC has defined 6 classes of risk: 1. Digital object format, 2. Software, 3. Hardware, 4. Associated organizations, 5. Digital archive, 6. Migration and derivative-based preservation plans. The 13 risks that Stanescue listed are:

- Whether or not royalties or license fees are or may be requested
- Whether the source or specifications can be independently inspected
- Whether revisions have maintained support for backward compatibility
- Whether it is complex or poorly documented
- Whether it is widely accepted or simply a niche format
- Whether competing or similar formats or components exist
- Whether embedded metadata can be mapped to other formats
- Whether DRM, encryption or digital signatures can be used
- Whether applicable expertise can be easily found
- Whether revisions happen so fast that the archive cannot keep up with demand
- Whether extensions, such as executable sections or narrowly supported features, can be added
- Whether authenticity can be easily compromised during transformations, either accidentally or maliciously
- Whether the associated organization or community is too small, in danger of collapsing, unique in its class or not easily replaceable

According to the INFORM methodology, the probability of each of these factors occurring will be measured by a 5-point scale, where 1 represents a low probability (less than 1% chance) and 5 is a high probability (more than 29%). The impact of this risk is measured on a 5-point scale from A-E, where A represents a minor risk and E represents a catastrophic risk. The two ratings are combined for a resulting risk exposure rating. The INFORM rating system is designed to be implemented by independent reviewers through several points over time, collated, and reviewed for trends and changes that may necessitate preservation actions.

As part of the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2 Project, Evelyn Peters McLellan produced a report on "Selecting Digital File Formats for Long-Term Preservation" (InterPARES, 2007). McLellan identifies five criteria to evaluate whether or not a file format is suitable for inclusion in a digital repository. These factors are: *widespread use*, *non-proprietary origin*, *availability of specifications*, *platform independence* (interoperability), and *compression*.

According to the InterPARES report, eighteen of twenty-four institutions reviewed considered widespread use of a file format to be an important criterion to inform the selection of file formats for their collections. The institutional interviewees indicated that there is a strong positive correlation between the widespread adoption of a file format and its continued support by the software industry. Similarly, file formats with a non-proprietary origin are ideal because non-proprietary formats are less dependent on software industry support; meaning, if the software industry ceases to support a non-proprietary file format, it will be easier for the community at large to create and adopt alternative methods to render the file.

In line with the criteria for non-proprietary file formats is the criterion that calls for the availability of documentation or specifications. If a file format has specifications available that can be used to create or re-create rendering software, it is considered to be less risky to collect items of that file format to preserve over the long-term. The criterion for platform independence is related to the degree to which a file format is dependent on particular hardware or software platforms. The less dependent a file format is, the more platforms it may be rendered on and the more it will be accessible over time. The last criterion, *compression*, involves whether or not and the degree to which a particular file has been compressed. Unlike the other criteria discussed, the *compression* criterion addresses the properties of a particular file and not those of an overarching file format. This approach is suitable for a list of inclusion criteria, though may not be applicable in general risk assessment processes.

In 2010, the Koninklijke Bibliotheek – the National Library of the Netherlands – published a report on the "Evaluating File Formats for Long-term Preservation," in which they examined seven criteria to be used in assessing file formats for long-term preservation. For each of the criteria examined, they included sub-criteria that, in conjunction with the primary criteria, were built into a weighted scoring table.

The first of these criteria is *openness*, as it relates to the ease of which one has access to information such as documentation that can be used to build or rebuild rendering software associated with a file format. The sub-criteria for *openness* are *standardisation, restrictions on the interpretation of the file format*, and *Reader with freely available source.* The *openness* criteria are very similar to the *non-proprietary* and *availability of specifications* criteria of the InterPARES report (2007), the Transparency factor from Arms and Fleischauer (2005), *openness* from the Kongelige Bibliotek (2004). The second criterion, *adoption* refers to how popular and ubiquitous the file format is. Sub-criteria for *adoption* are, *worldwide usage* and *usage in the cultural heritage sector as archival format.* The *complexity* criterion has the sub-criteria, *human readability, compression,* and *variety of features.* These criteria are related to the notion of how much effort has to be put into rendering and understanding the contents of a particular file format.

The *technical protection mechanism* criterion, like the *compression* criterion, addresses characteristics of an individual file and not a file format in general. This criterion includes characteristics that prohibit access to a file's contents: password protection, copy protection, digital signature, printing protection, and content extraction protection. The *self-documentation* criterion includes the sub-criteria of *metadata* and *technical description of format embedded*, and refers to the degree to which information that may be useful in accessing the contents of a file are able to be included in or associated with the file itself.

Robustness refers to how vulnerable the file format is to corruption or becoming unrenderable as a result of technological and environmental changes over time. Sub-criteria include: *robust against single point of failure*, *support for file corruption detection*, *file format stability*, *backward compatibility*, and *forward compatibility*. The final criterion discussed is *dependencies*, which refers to whether or not a file format is dependent on particular hardware or software platforms to be accessible. Sub-criteria included under *dependencies* include: *not dependent on specific hardware, not dependent on specific operating systems, not dependent on one specific reader,* and *not dependent on other external resources*.

As part of the MathDiss International and the EMANI projects that gather and use mathematic research that is created using the LaTeX format, researchers evaluated the LaTeX format as a suitable archival format. During this project, the project teams established and evaluated criteria for archival file formats (Fischer, 2003). These criteria included: *error tolerance*, *long-term stability*, *full open specification*, *system dependence*, *ease of handling*, *independence of commercial interests and influence*.

Error tolerance refers to the degree to which a file format can tolerate bit corruption without becoming unreadable. *Long-term stability* is related to how many versions a file format exists in over time. If a file format is continually changing versions, and therefore not stable, it will be more difficult to find software that can render files in that format. *Full open specification* refers to whether or not the specification of the file format is publicly available for use in recreating software to create and/or access content stored in that format. System independence, as describe in the paper, is related to whether or not a file format is dependent on a particular hardware platform, but one might also take this to mean software system dependence as well.

The criterion is unique among the lists of criteria discussed here in that it explores how easily a file format may be handled by a system and the ease with which it may be transformed into a more simply handled format. *Ease of handling* refers to the computing capacity required to access and handle the format as well as the complexity or number of files that are necessary to assemble in accessing the content of a particular file format. The criterion for *independence of commercial interests and influence* relates to whether or not commercial enterprises may prevent or inhibit present or future access to file formats under their purview.

The UK National Archives contracted out the writing of two reports on file formats: “Selection of Preservation Formats: Trends and Issues,” and “Criteria for the Selection of Preservation Formats” (2005a; 2005b). The latter contains a detailed description of categories and lists of criteria to consider when selecting file format for preservation purposes. Twenty criteria are suggested and are grouped under four categories: *popularity*, *utility*, *technical features*, and *commercial factors*.

Under the *popularity* category are the criteria *ubiquity, numbers of viewers available, disclosure, and stability and backwards compatibility*. *Ubiquity* refers to the extent to which the format is in use. *Numbers of viewers available* refers to the number of different software products available to render the file and *disclosure* is the degree to which the format specifications are publicly available. *Stability and backwards compatibility* refers to both the speed with which the file format software is upgraded and its ability of newer versions of the software to render files created in older versions of the software.

The criteria related to *utility* refer loosely to certain aspects of the file format that increase its likelihood of being renderable and useful in the future. These criteria are *loss of "functionality" caused by migration*, r*etention of metadata on migration, proportion of digital objects/records,* and *support for redaction.* All of these criteria are self-explanatory except for *proportion of digital object/records,* which refers to the number of digital objects or records that are stored in a particular file format. According to the report, the more records that are stored in that format, the more important and consequently the more preservable the file format will be.

The *technical features* criteria are *ease of identification, technical dependencies, transparency, probability of information loss due to corruption, likelihood of infection, ease of validation, availability of migration path,* and *losslessness. Ease of identification* is the degree to which a file format can be identified using available tools such as automated file format identification software. *Technical dependencies* is the extent to which the file format is dependent on hardware or software platforms. Unlike other metrics discussed, the *transparency* criterion does not refer to open disclosure of file format specifications, but rather the degree to which access to the file is inhibited by technical limitations such as encryption, compression, and digital signatures. Compression is also considered in the *losslessness* criterion, which covers compression or any other activity that may cause loss of information in the file. *Ease of validation* covers the ease with which deviations from the file format specifications can be detected. *Availability of migration path* refers to the availability of new file formats that the file format may be migrated to if necessary.

Finally, the *commercial factors* category involves factors that are related to commercial entities. The factors included in this category are *cost of migration to the format, cost of migration away from the format, IPR impacts,* and *verbosity.* According to the report, the *cost of migration away from the format* and *cost of migration to the format* factors involve the costs of migrating from one file format to another. Costs here are related to human effort, software license fees, the necessity to use dedicated hardware, and preparation of storage. *IPR impacts* refers to the Intellectual Property Rights (IPR) and how they may inhibit the access to information in a file. *Verbosity* refers to the number of bytes required to store information in a particular file format. The more bytes that are required to store a file, the more expensive storage will be in the long term.

In a report on file formats for preservation in digital libraries, Sunita Barve described challenges in file formats (2007). In this report, she describes several factors that could inhibit continued access to information encoded in particular file formats. The following is an abbreviation of Barve's list:

- Developer of file format goes out of business
- Developer stops supporting that format
- The market share of the developer declines
- Supporting program of the software change significantly
- Third party support is lacking
- Format depends on obsolete hardware or operating system
- Format is proprietary
- New versions of application software may not support earlier format versions
- Application software developers do not release documentation or release inaccurate documentation
- Software too complex and document structure layout is proprietary

The *Groupe Pérennisation des Informations Numériques* (PIN) or the Sustainability of Digital Information Group of France produced a report on criteria for evaluating file formats for long-term preservation (Huc et al, 2004). In this report, they discuss one underlying condition, three primary criteria, and five additional criteria for evaluating a file format's suitability for long-term preservation.

The underlying condition is that a, "data format can only be acceptable for information preservation if it is fully known to the entity (organization) in charge of the preservation" (p. 1). The format is "known" only if it is in an open format that is public and copyright free, or if it is in a format that has available an "exhaustive and validated description" (p. 6).

The first of the three main criteria is a format's capability to represent all the information that it is meant to preserve. The second main criterion is that the file format should be publicly standardized. This criterion recommends the use of both standardized and non-proprietary file formats. The third main criterion is titled, "Possibility of modifying the data at a later date" and has the description, "When the need to be able to modify a document is a factor, the format must be selected accordingly" (p. 8). The supporting text indicates a need to preserve the authenticity of a document over time, so though it is not explicitly stated, one could assume that file format selection should either not allow changes to be made or it should allow for changes, but be able to maintain a document's authenticity in some way. Overall, the intention of this third main criterion is not very clear.

The first of the additional criteria is *tools and creation and manipulation facilities* which means that one should consider the cost and availability of tools needed to create, access, and manipulate the information in a particular file format. The second consideration is the *complexity /simplicity of the format* where a simpler format is preferred over a complex one. The third additional criterion is *inspectability,* or the ability to automatically verify that the format complies with format specifications and rules and restrictions established for preservation. *Metadata* is the fourth criterion and relates to the ability to automatically extract metadata from the file. The last criterion is *performance of available implementations*, which relates to the ability to or the availability of analysis of file format performance information. Performance analysis information should be used to assess whether or not a file format is suitable for long-term preservation.

As part of the Preservation and Long-term Access through Networked Services (PLANETS), Christoph Becker drafted a report on preservation planning services provided in the Plato 2 tool (2008). The report describes, as part of the risk assessment service in Plato 2, a list of generic risk factors and a scoring rubric considered for use in the tool. The ten risk factors and their definitions are listed in *Table 2.4.2*. Each of these factors has three “allowed values” that each have associated with them numeric scores that are used to rate a file format’s level of risk.

In 2000, the Council on Library and Information Resources (CLIR) funded the Risk Management for Digital Information Project for which seven categories were created for risks associated with file formats-based migration for image collections (Lawrence, et al, 2000; Rieger & Kenney, 2000). Though these risk categories were created in relation to file format migration of images, the categories are largely similar to those discussed throughout this paper. These categories are *content fixity*, *security*, *context and integrity references*, *cost*, *staffing*, *functionality*, and *legal*.

Risk Factor	Definition
Ubiquity	The degree of adoption of the format
Support	The number of access tools currently available
Disclosure	The extent to which the format documentation is publicly disclosed
Document Quality	The accuracy and completeness of the available documentation
Stability	The speed and backward compatibility of format changes
Ease of Identification	The ease with which the format can be automatically identified
Ease of Validation	The ease with which the format can be automatically validated
Use of Compression	The nature of any compression used
Intellectual Property Rights (IPR)	The extent to which the format is encumbered by IPR issues
Complexity	The degree of content and behavioral complexity supported

Table 2.4.2. File format risk factors created for the Plato 2 tool. (Becker, 2008, p. 9)

Content fixity refers to the level of the risk that the bit configuration of a file in a particular file format can be altered. *Security* refers to changes in security measures such as watermarks and digital stamps that can be affected by migration. *Context and integrity* refers primarily to the contextual relationships that a particular file has with other files, and with other software and hardware platforms. Similarly the risk category, *references,* refers to contextual relationships that are created by links and images. *Cost* refers to the long-term costs associated with migrating a file to a new file format. *Staffing* refers to the risks associated with maintaining the necessary staff to preserve a migrated file over time. The *functionality* category refers to the risks associated with loosing functionalities such as printing, display variations, and interface modifications when a file is migrated to a new format. The *legal* risk category refers to legal complications such as copyright regulations that may prevent or be affected by migrating a file to a new format.

Researchers at the University of Minho in Portugal proposed a Service-Oriented Architecture for automatic migration of file formats (Ferreira, Baptista, & Ramalho, 2006; 2007). One of the elements of the proposed architecture is a "Format Evaluator" that provides information about file formats. The Format Evaluator was designed to consider a list of criteria to determine potential benefits of migrating a digital object from one file format to another. These criteria are listed in *Table 2.4.3*.

Criterion	**Description**
Market share	The degree to which the file format has been adopted
Support level	Whether the creator continues to provide technical support for the format
Is standard	Whether or not the format is considered a standard by standards organizations
Open specification	Whether or not the format specifications can be inspected
Supports compression	Whether or not the format allows compression
Lossy compression only	Whether or not the file format only allows lossy compression
Supports transparency	Whether or not the format supports transparency. The term transparency here refers to visual transparency such as is supported by raster images.
Embedded metadata	Whether or not the format can accommodate embedded

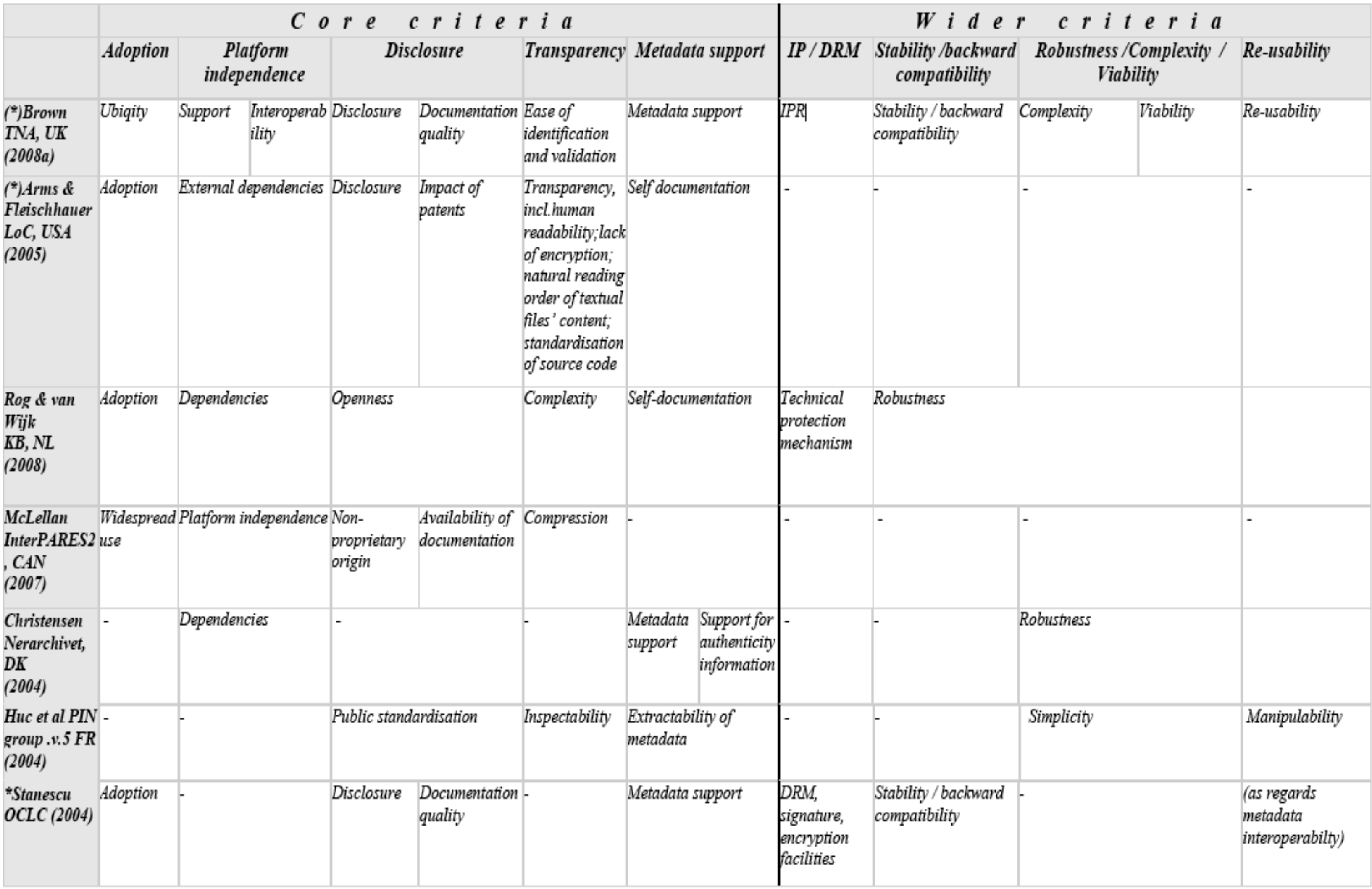
Criterion	Description
	metadata
Royalty-free	Whether or not the format requires royalty fees to use or produce the format
Open source	Whether or not there is open source software available that can render files saved in the format
Backwardly compatible	Whether or not newer versions of the rendering software can render files from older versions
Documentation level	Whether the format specification is well documented
Competing formats available	Whether there are similar formats available
Digital Rights Management (DRM) support	Whether DRM measures can be used on the format
Update frequency	How frequently the file format is updated into new versions
Supports custom extensions	Whether the format supports custom extensions such as executable code or narrowly supported features
Life time	The length of time a file format has existed
Transparent encoding	The degree to which the format can be read using simple tools such as a text editor
Reader single producer	Whether or not the reader is produced by only one entity
Single reader	Whether or not there is only one source of software to render the file
Open source reader	Whether or not the source-code of the rendering software is publicly available
Multiplatform reader	Whether or not the rendering software can run on multiple platforms

Table 2.4.3. Format Evaluator criteria (Ferreira, Baptista, & Ramalho, 2006)

Many of the evaluation factors discussed here overlap: the lists either contain the same terms or contain terms that convey the same meaning as terms in other lists. Representatives from the Digital Curation Centre, Digital Preservation Coalition and the UK National Archives have produced literature in which they analyzed the criteria from several of the lists of criteria discussed above.

Stephen Abrams (2007) wrote an installment on "File Formats" for the Digital Curation Centre's Digital Curation Manual series in which he described the file format assessment criteria presented by Stanescu for the INvestigation of Formats based on Risk Management (INFORM) project (2004), Huc for the Groupe Pérennisation des Informations Numériques (PIN) (2004), Arms and Fleischhauer for the Library of Congress (2005), Brown for The National Archives (TNA) of the UK (2003), and Christensen for the Kongelige Bibliotek (Danish Royal Library) (2004).

Similarly, the Digital Preservation Coalition published a Technology Watch report by Malcom Todd on "File Formats for Preservation" (2009). The report contains a chart of file format evaluation criteria that were covered by the 2007 DCC report, with the addition of criteria presented by Rog and van Wijk of the Koninklijke Bibliotheek (KB – Royal Library of the Netherlands) (2008), and McLellan for the InterPARES2 project (2007).



	Core criteria								Wider criteria				
	Adoption	Platform independence		Disclosure		Transparency	Metadata support		IP / DRM	Stability /backward compatibility	Robustness /Complexity / Viability		Re-usability
(*)Brown TNA, UK (2008a)	Ubiqity	Support	Interoperability	Disclosure	Documentation quality	Ease of identification and validation	Metadata support		IPR	Stability / backward compatibility	Complexity	Viability	Re-usability
(*)Arms & Fleischhauer LoC, USA (2005)	Adoption	External dependencies		Disclosure	Impact of patents	Transparency, incl.human readability;lack of encryption; natural reading order of textual files' content; standardisation of source code	Self documentation		-	-	-		-
Rog & van Wijk KB, NL (2008)	Adoption	Dependencies		Openness		Complexity	Self-documentation		Technical protection mechanism	Robustness			
McLellan InterPARES2 , CAN (2007)	Widespread use	Platform independence		Non-proprietary origin	Availability of documentation	Compression	-		-	-	-		-
Christensen Nerarchivet, DK (2004)	-	Dependencies		-		-	Metadata support	Support for authenticity information	-	-	Robustness		
Huc et al PIN group .v.5 FR (2004)	-	-		Public standardisation		Inspectability	Extractability of metadata		-	-	Simplicity		Manipulability
*Stanescu OCLC (2004)	Adoption	-		Disclosure	Documentation quality	-	Metadata support		DRM, signature, encryption facilities	Stability / backward compatibility	-		(as regards metadata interoperabilty)

Figure 2.4.1. Table of core and wider criteria (Digital Preservation Coalition, 2009)

In *Figure 2.4.1.*, Todd aggregated the criteria from the seven sources into five core and four wider criteria. The core criteria are *adoption*, *platform independence*, *disclosure*, *transparency*, and *metadata support*. The wider criteria are *IPR/DRM*, *stability/backward compatibility*, *robustness/complexity/viability*, and *re-usability*.

According to the glossary included in the report, *adoption* is the extent to which a format is in use. *Platform independence* is the extent to which the format is supported by hardware and software platforms. *Disclosure* is the extent to which a format specification is in the public domain. *Transparency* is the ability of a format to be inspected in order to discover its identity. *Metadata support* is the ability of a format to include metadata such as representation information. *IPR/DRM* is the extent to which a format supports/allows for restrictions related to intellectual property rights and digital rights management. *Stability/backward compatibility* is the extent to which a format is subject to version increases and is able to be rendered by newer software versions. *Robustness/complexity/viability* is a file format's resistance to corruption. *Re-usability* is the extent to which a file format can be accessed and reused.

Preceding both of these reports are two 2005 reports commissioned from Cornwell Management Consultants by The National Archives of the UK. The first is a synthesis of existing file format criteria (The National Archives, 2005a; 2005b). Cornwell Management Consultants led a process involving analysis of 90 criteria that had been included in previous file format criteria evaluation lists, a workshop to review and refine these criteria, and an ensuing iterative process of telephone interviews and further criteria refinement. Through this process, 48 of the 90 criteria were mapped into the 20 criteria for file format evaluation previously described.

It is important to note that existing literature focuses on criteria for the selection of file formats as sustainable vehicles for the long-term preservation of information, i.e., which formats are the best to include in an archival digital repository. The focus of my proposed research is somewhat different. Instead of determining whether a file format will be suitable for preservation purposes in an archival digital collection, the factors being considered here are being evaluated on whether or not they cause or are formative indicators of file format endangerment. In order for a system to notify individuals of potential file format endangerment, a suitable metric must be in place. An assumption in this proposal is that a memory institution may be given the responsibility of managing *any* file format, and a file format endangerment index should be designed to measure the endangerment of all formats, not just whether a format is suitable for inclusion in the collection.

The criteria previously created and discussed are a valuable start to the development of a comprehensive index that can be used to assess file format endangerment. The methodology used to create such an index can and should be informed by the examination and assessment of these previously created measures.

2.5. Parallel Problems in Conservation Biology

I have identified the gap in current digital preservation practice wherein there is presently no applied file format risk monitoring and analysis method. In my exploration of methods to best address this gap, I discovered similar problems in the research area of conservation biology. After further exploration, it was clear that the methods used to monitor, analyze, and warn against impending species extinction are relevant to my research in file format endangerment.

Biologists and allied researchers have invested enormous time and effort to identify, categorize, and count the organisms of the Earth. Consequently, many methods for monitoring species have been created, tested and refined. There is much relevant research in this area and there are several possibilities to adapt these methods for use in monitoring file format endangerment.

At the heart of species endangerment monitoring is population viability analysis (PVA). Doak defines PVA as, "the use of quantitative methods to predict the likely future status of populations of conservation concern and also to predict how best to manage these populations" (2009, p. 522). While there are many variations of PVA, the method generally involves the collection of data for specified factors that inform an understanding of a species' health in the wild. The data is then statistically analyzed, often using analysis and simulation software to predict the possibility of a species' extinction. The lists of factors vary depending on the species and have evolved over time as a product of increased understanding of what kind of data can be collected and what does and does not affect the predictability of species endangerment.

The roots of PVA are found in the work of Shaffer (1981) and Soulé (1985). Shaffer's work in minimum population sizes helped shape models for predicting the possibility of extinction that later became part of contemporary PVA models. Soulé has been instrumental in shaping and defining the field of conservation biology, the field in which researchers most commonly use PVA methods.

In their overview of PVA, Gerber and González-Suárez wrote that,

> PVA represents one of the most valuable approaches that has emerged from the burgeoning field of conservation biology. While it is impossible to make precise predictions about the exact time to extinction, PVA has offered useful tools to

> estimate the relative risk of extinction, and to compare the efficacy of alternate management strategies (2010, Summary, ¶1).

Applying a PVA-type approach to file format endangerment analysis may not be able to predict an exact date when a file format will become endangered, but it can be used to alert the community to potential file format endangerment. Additionally, creating PVA-style simulations may be useful in comparing file format preservation strategies.

In 1995, Paul Angermeir reported on a study in which he examined the attributes of extinction-prone species of freshwater fish in Virginia. Based on the results of this study, Angermeir was able to make focused suggestions on the prevention of extirpation and broader extinction for these fish species. The method Angermeir used involves analyzing data collected for individual factors of extirpated species. It was through correlation analysis of these factors between species that he was able to determine the similarities between the extirpated species. He stated that, "because direct observation and experimentation are generally infeasible, correlative analyses are the primary tools available to study large-scale extinction processes" (p. 154). Angermeir acknowledged the difficulty in detecting patterns in systems with complex dynamics. In order to overcome the resulting "statistical noise," he performed multiple complimentary analyses. Because the "ecosystem" of file formats is also complex, similar triangulation of methods may be useful in analyzing factors that contribute to file format endangerment.

The notion of extirpation, or the local extinction of species, is useful in considering research methods for file format endangerment. By extension, the phrase "file format extirpation" would mean that a local institution or its regular users could not access information stored in a particular file format. Angermeir stated that, "extinction is rarely cataclysmic. Rather, it is incremental, with total extinction preceded by local or regional extinctions" (p. 144). He said that knowledge of local extinction of a species could help to fuel proactive measures to prevent further extirpation and widespread extinction. Understanding the causes of local extinction can help in the creation of more specific solutions to the problem. By extension, understanding localized file format endangerment can be useful in creating more useful solutions, both locally and worldwide.

Mace and Lande (1991) provided some guidelines for designing an effective extinction threat assessment system. They suggested the following six characteristics of an ideal system:

- The system should be simple. There should only be a few categories for assessing risk, they should have a clear relationship with each other, and "should be based around a probabilistic assessment of extinction risk"
- The categorization system should be flexible about the quality and quantity of data required.
- The system should work with any species.
- The terminology used should be clear.
- The system should include some assessment of uncertainty.
- A timescale should be used for each category of extinction, i.e., number of years until extinction. (p. 150).

All six of these characteristics of an ideal extinction threat assessment system could be relevant to the development of an ideal file format endangerment assessment system.

O'Grady, Reed, Brook, and Frankham (2004) examined the correlation between sixteen criteria used in determining species extinction risk. They performed stepwise multiple regression analysis on each of the factors and found that, "population size and percent change in population size are the best predictors of extinction risk" (p. 519). O'Grady, Reed, Brook and Frankham measured sixteen parameters, and of the sixteen parameters measured, the best predictors of extinction risk were population size and change in population size. This is significant in that O'Grady and his colleagues showed that monitoring only the population size of a species is sufficient to predict impending extinction. Once file format endangerment monitoring factors have been selected and sufficient data has been collected, similar tests of correlation should be performed to assess effectiveness. If the models are truly similar, it could mean that monitoring the number of instances of a file format and the change in this number may be sufficient for effective endangerment prediction.

2.5.2. Applications in File Format Endangerment Research

The research area of conservation biology provides a strong foundation and a useful framework from which to base file format endangerment research. The conservation biology field presents decades of research in methodologies to track and preemptively detect threats to the continued existence of living species. Examining the frameworks and methods used in conservation biology has helped me to define the steps that need to be taken to effectively assess file format endangerment.

First, conservation biology has tried and tested methods for collecting and analyzing data for monitoring threats in complex systems. In particular, the methods of Population Viability Analysis have been used, tested, and improved and provide a strong foundation from which to base file format endangerment analysis. Second, conservation biology presents a useful framework of threat evaluation and terminology, some of which I have appropriated for this research. In particular, I am using the term "endangerment" to refer to the possibility that information encoded in a particular file format will become inaccessible within a certain timeframe; i.e., in 20 years or more. This definition also reflects my adherence to Mace and

Lande's (1991) recommendations to base threat systems on "probabilistic assessment of risk" and to include a timescale in each category of risk. Third, the results presented in O'Grady, Reed, Brook, and Franklin (2004) served as a basis and strong motivator for disambiguating and reducing the wide array of file format evaluation factors discussed in the literature to only the most relevant.

2.6. Formative Indicators and Index Construction

One of the common elements between conservation biology and file format endangerment methods is the collection and analysis of data for pre-defined factors to detect potential dangers. The pre-defined factors represent indicators of the phenomenon being measured, i.e., species endangerment, epidemics, or file format endangerment; and are commonly called *formative indicators*.

Formative indicators, used in index construction, have an opposite relationship than do "effect" or "reflective indicators," which are commonly used in scale development. In *Figure 2.6.1,* the opposite causal directions of reflective and formative measurement models are illustrated, where η is the construct or phenomenon being measured, and x_1, x_2, and x_3 are the reflective and formative indicators. In panel 1, λ represents the relationship that the construct has on the reflective indicators, x_1, x_2, and x_3. The symbol ε represents the error. In panel 2, ζ is an error or disturbance term that represents remaining relationships of the construct that are not represented by the formative indicators and that cannot be measured. The symbol γ represents the relationship that the formative indicators, x_1, x_2, and x_3 have on the construct and the *r* variables and their incumbent arrows represent their interdependency toward defining, creating, and causing the construct.

As an example of a formative measure, the construct or the phenomenon that I intend to measure is file format endangerment, and the formative indicators are the factors that are determined to be causes of file format endangerment. In a reflective measure, the effects, i.e. the reflective indicators of the phenomenon, are measured, such as in personality measures where the personality is the construct and the personality traits are measured as an effect of the personality. According to Bollen, “most researchers in the social sciences assume that indicators are effect indicators,” where, “cause indicators are neglected despite their appropriateness in many instances” (1989, p. 65).

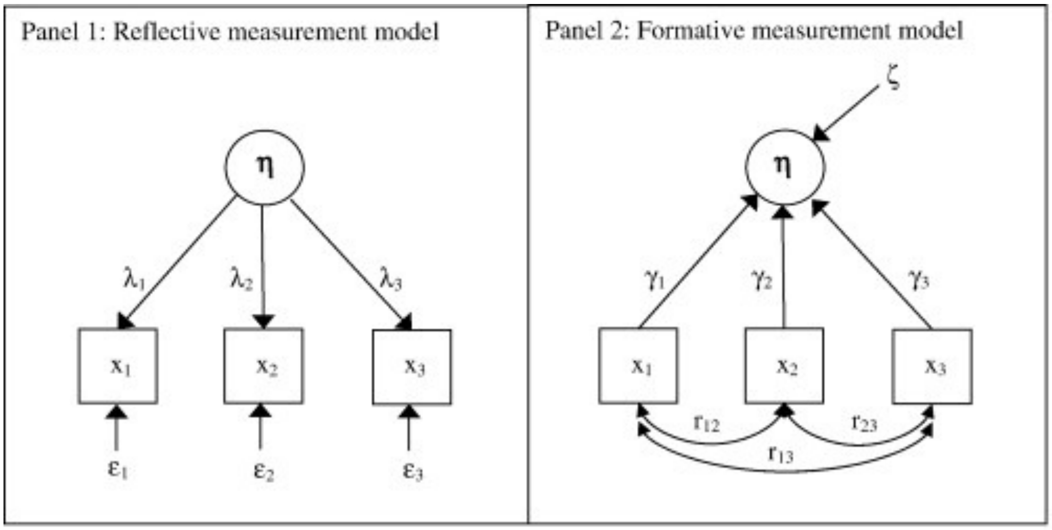

Panel 1: Reflective measurement model
Panel 2: Formative measurement model
η
λ1
λ2
λ3
x1
x2
x3
ε1
ε2
ε3
ζ
η
γ1
γ2
γ3
x1
x2
x3
r12
r23
r13

Figure 2.6.1. Causal direction in reflective and formative measurement models (Diamantopoulos, Riefler, & Roth, 2008, p. 1205).

It is often not obvious which of the two measurement models is most appropriate. Bollen (1989) suggests that one method of determining which model is more appropriate is to perform a “temporal priority” mental experiment, or simply put, think about which happens first: the indicator or the construct. In the case of file format endangerment, my intention in this research is to create a predictive model using factors that precede endangerment. Consequently, such a model demonstrates the temporal priority of factors that are exhibited before the phenomenon of file format endangerment. Phenomenon prediction requires data collection for *a priori* factors, or observable factors that occur before the measured phenomenon; therefore, a formative measurement model best suits the purposes of evaluating the possibility that information encoded in a particular file format will become inaccessible within a certain timeframe.

Once a researcher has determined that the indicators in question have a formative relationship with the construct, they can begin to design the measurement model, or index. Diamantopoulos and Winklehofer (2001) describe the four steps for constructing an index:

1. Content Specification - defining the "domain of content the index is intended to capture" (p. 271).
2. Indicator Specification - choosing the indicators to be added to and tested for the index.
3. Indicator Collinearity - checking that there is not excessive collinearity between the indicators.
4. External Validity - determining that the index measures what it claims to measure and "assessing the suitability of the indicators" (p. 272).

Diamantopoulos and Winklehofer suggest that the definition of the domain be broad enough to encompass all of the causal indicators. Though they provide no formal recommendation for specifying which indicators to include in an index, they reported that they selected indicators for their export market sales forecasting index through "an extensive review of the forecasting literature as well as exploratory interviews with export managers" (p. 272). Selecting the right indicators for a formative measure is very important since the construct being measured is defined by the indicators. Consequently, "changes in the measures [indicators] are hypothesized to cause changes in the underlying construct" (Jarvis, MacKenzie, & Podsakoff, 2003).

In respect to indicator collinearity, formative indictors in indexes should have a direct effect on the phenomenon being measured and have little to no intercorrelation, meaning the indicators in a formative measure should have little to no direct effect on each other. While indicators in a formative measure may have some interaction with each other, it is best if they do not have strong correlations with one another (Petter, Straub, & Rai, 2007).

Finally, determining external validity involves testing the index to determine if it measures the specified construct. Diamantopoulos and Winklehofer suggest, "One possibility is to use as an external criterion a global item that summarizes the essence of the construct that the index purports to measure" (p. 272). Another suggestion is to include some reflective indicators in the model as a secondary measure of the appropriateness of the formative indicators, in what is called a multiple indicators and multiple causes (MIMIC) model (Bollen, 1989). In the MIMIC model, shown in *Figure 2.6.2.*, the construct η is caused by the formative indicators, x_1, x_2, and x_3, and in turn causes the reflective indicators y_1 and y_2. According to Bollen, measuring resulting effects of the construct in conjunction with the causes demonstrates the model measures what it purports to measure.

The research presented here addresses the first two of the above steps. For the first step, I specify the content of the file format endangerment index as being all factors that cause, either through their presence or absence, information encoded in particular file formats to become inaccessible over a specified timeframe. Similarly to Diamantopoulos and Winklehofer, I addressed indicator specification through an extensive literature review, supplemented by the factor-rating Delphi exercise. I intend to address steps three and four in future research.

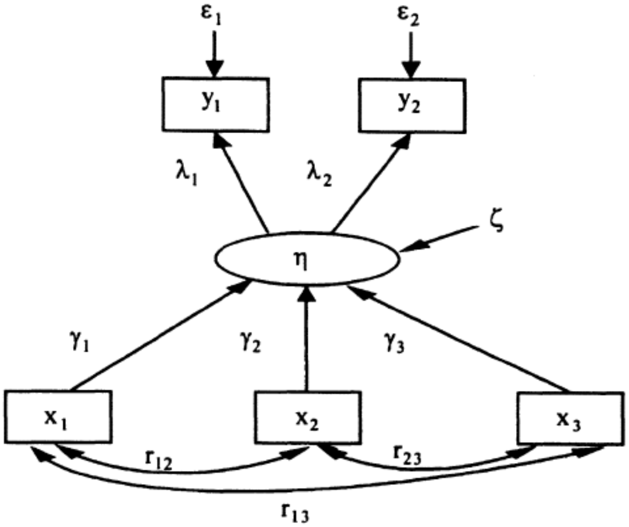


Figure 2.6.2. Multiple indicators and multiple causes (MIMIC) model (Diamantopoulos & Winklehofer, 2001, p. 272).

CHAPTER 3: RESEARCH DESIGN AND METHODS

As described in the literature review, there has been a clear call for file format risk assessment from the digital preservation and memory institution community. There have been several attempts to answer this call, but unfortunately there are several impediments to this being achieved. The first impediment is the lack of understanding of which factors truly cause the risk. The second is the lack of sufficient data to reliably estimate file format risk. The third impediment is the lack of baseline knowledge of current risk from which to compare future measurements. The fourth impediment is the lack of tested methods to collect and evaluate file format risk.

My review of the literature revealed a number of initiatives that either attempted to build file format risk assessment functionality into a system or attempted to outline file format evaluation factors. A close examination of the dozens of factors described in the literature reveals a diversity of purpose and application. While the existing lists of factors are a good start at exploring the question of what causes a file format to be a less viable means of encoding and retrieving digital information over time, they fail to provide a method to truly assess risk.

A promising next step toward assessing file format endangerment is to examine each of the dozens of factors discussed in the literature to determine which are direct causes of endangerment. From there, these factors can be operationalized as formative indicators in a file format endangerment index. This approach is based on research reported by

Diamantopoulos and Winklhofer (2001), who suggest that an extensive literature review and exploratory interviews with managers as experts is a viable method for choosing an initial set of index factors.

The primary objectives of this research are to establish a baseline understanding of current file format endangerment levels, and to clarify which of the many factors discussed in the literature are the most relevant formative indicators to include in a file format endangerment index. The research described here took a three-pronged approach to addressing these issues: two separate Delphi studies and one information gathering and rating exercise designed to test a unification of the two Delphi studies.

The Delphi method was the most effective method to establish a baseline file format endangerment level and to determine which are the most relevant causal factors of file format endangerment. When little data exists on a topic, such as with file format endangerment, Delphi is known to be an effective method of “producing trustworthy personal probabilities regarding hypotheses” in experts’ knowledge area (Helmer & Rescher, 1959, p. 38). Dalkey (1968) explained that characteristics of a Delphi procedure are anonymity, iteration with controlled feedback, and statistical group response. These procedures were designed to reduce “the influence of certain psychological factors, such as specious persuasion, the unwillingness to abandon publicly expressed opinions, and the bandwagon effect of majority opinion” (Gordon & Helmer, 1964, p. 5). Gordon and Helmer suggested that inviting participants to review other panel members’ reasoning will promote a thoughtful consideration of ideas and will lead to a more accurate representation of the truth. This process reduces what they call the “bandwagon effect,” or the propensity to join consensus based on social pressures. Several comparative studies demonstrated Delphi’s effectiveness in reducing bias and normative pressures (Boje & Murnighan, 1982; Stasser & Titus, 1985; Van de Ven & Delbecq, 1974).

After performing Bollen's (1989) temporal priority mental experiment, described in *Section 2.6* above, I determined that the factors I was examining for file format endangerment occurred before the phenomenon of file format endangerment. This pre-phenomenal occurrence indicates that the factors could be causes of file format endangerment, and thus appropriate for use in an index. Working toward construction of an index, rather than a scale, reflects the community need to gather specific, operationalizable information about causes of file format endangerment.

3.1. Research Questions

In order to better understand the nature of file format endangerment and the factors that cause it, I designed the research discussed in this dissertation to answer the following research questions:

1. Do digital preservation experts believe that certain file formats pose a risk for digital preservation?
2. Which file formats do digital preservation experts believe are endangered?
3. What are the most relevant formative indicators of file format endangerment, and how can these indicators be measured?
4. How effectively can the expert chosen file format endangerment factors be applied to rating file format endangerment?

3.2. Units of Analysis

This research examined two units of analysis: 1) digital file formats, and 2) factors to be included in a file format endangerment index. During the study I asked expert research participants to rate a set of fifty file formats on an endangerment scale. File format endangerment factors are the second unit of analysis as participants consider them individually and rate them on a scale of relevancy as a cause of file format endangerment. The highest-rated factors were tested by a trained reviewer who applied them in rating endangerment levels of a set of forty-three test file formats.

3.3 Definitions

For the purposes of this study, a file format is, "the internal structure and/or encoding of a file which allows it to be interpreted or rendered in human accessible form" (The National Archives, 2005, p. 8).

File format endangerment indicates the possibility that information stored in a particular file format will not be interpretable or renderable in human accessible form within a given timeframe. Early in the development of this research design, I defined file format endangerment as when a file format is in danger of not having software available to render information that is encoded in that format. Considering the fact that a number of factors discussed in the literature relate to the availability of rendering software, I realized that using a definition that centered solely on the availability of rendering software was problematic. To address this conflict, I changed the definition to be centered on the possibility that content encoded in a particular file format will not be interpretable or renderable. Inaccessible in this context means that information is not capable of being used or seen.

3.4. Research Design

This research involved the use of four questionnaires; administered online using Qualtrics survey software:

1. A questionnaire designed to collect information about the quantity and quality of experience that recruited Delphi participants had working with file formats in a digital preservation context. I used the information collected from this questionnaire to determine the expertise level of participants and to assign them to one of the two Delphi groups.
2. A questionnaire designed to collect information on participant opinions of file format endangerment level ratings of 50 test file formats. This questionnaire was designed to be used in a Delphi process in which participants answer the questionnaire over multiple rounds and review anonymous responses of their fellow participants between rounds. This questionnaire was designed to collect information on a baseline level of file format endangerment and to collect information about which factors participant consider when rating file formats.
3. A questionnaire designed to collect information on participant opinions of the relevance of factors as a cause of file format endangerment. This questionnaire was designed to be used in a Delphi process in which participants answer the questionnaire over multiple rounds and review anonymous responses of their fellow participants between rounds. This questionnaire was designed to clarify which of the many factors discussed in the literature are considered by experts to be direct causes of file format endangerment.
4. A questionnaire designed for one special rater participant to collect and report on information about factors for a list of file formats, to collect endangerment level ratings for the list of file formats, and to collect relevancy ratings for the list of factors considered as causes of file format endangerment. I designed this exercise to provide an additional source of data collection for both understanding the current perceived level of file format endangerment and for understanding which factors are direct causes of file format endangerment. This was also designed to address Research Question 4 by bringing together the two pieces of the study: file format rating and factor rating together and to test how they function as a whole.

These questionnaires were administered in the pilot testing and in the final research design.

The final research design is comprised of the following steps, described in detail below:

1. Select file formats and factors
2. Recruit participants for the two Delphi studies and factor testing study
3. Administer Questionnaire 1
4. Administer Questionnaires 2 and 3, Round 1, in tandem
5. Administer Questionnaires 2 and 3, Round 2, in tandem
6. Calculate Spearman's … coefficients of Round 1 and Round 2 results
7. Continue to additional round as necessary

8. Administer Questionnaire 4
9. Administer special rater follow-up interview
10. Analyze data

3.5. Selecting File Formats and File Format Endangerment Factors

A crucial first step in conducting this research was selecting the file formats and file format endangerment factors that participants would rate. I selected through a process of elimination from an original list of one hundred file formats from a list of file extensions collected by the U.S. National Archives and Records Administration (NARA). I selected the factors to be rated through a process of compiling several file format evaluation factor lists from the literature. I describe both of these processes in detail below.

3.5.1. Selecting File Formats

I selected the file formats from a list of file extensions that were collected by the National Archives and Records Administration's (NARA). The original dataset I reviewed contained a list of every file extension and a count of instances of the file extension that appeared in their digital collections. Having access to this dataset presented me with a list of formats that are in a real-world digital collection. This data was collected in April of 2012 as part of the CyberInfrastructure for Billions of Electronic Records (CI-BER) cooperative research agreement between NARA, the National Science Foundation, and the University of North Carolina at Chapel Hill (Sustainable Archives and Leveraging Technologies group, 2010).

The file format extension data represents record groups from all contributing U.S. federal agencies, excluding NASA. I only had access to the file extension data and did not have information on the version the extensions referenced. I searched for information online to help me identify the file formats associated with each extension. Out of 1,497 distinct file extensions represented, I chose 100, based on the most frequently appearing extensions, and after disqualifying a number of extensions that:

- *I could not identify.* There were several extensions for which I could find no information that helped me to identify the file format. For example, I could not identify the extensions: .nh2, .SOIL, .inv, or .hei.
- *Were not supplemental files or formats.* A number of file extensions were peripheral files to core file formats. For example, ArcGIS, a graphical information system, produces a number of peripheral formats represented by file extensions in the NARA list such as .rrd, .freelist, .atx, and .mxd.
- *Were not redundant extensions of the same file format*. There were several instances of file extensions that represented the same type of format, for example: .tif/.tiff, .htm/.html, and .jpg/.jpeg. These extension names and extension counts were combined in the final list.
- *Were not compression or aggregation formats.* I removed the extensions that were obviously compression or aggregation formats like .zip, and .tar, in order to focus this research effort on simple formats. I did miss one compression format, .kmz, which ended up in the final list.
- *Were not generic extensions that represent several file formats.* There were several extensions that did not clearly refer to one file format, and were most likely user-named or computer generated extensions. Examples of these are .rpt which is used by a variety of applications to identify report files, and .bak which is a common extension used for file backups.
- *Were not changed by NARA employees*. For example, files with the extension .corr01 represent files that were corrected one time by NARA employees, .corr02 represent files that were corrected twice, and so on. After making corrections to these files, NARA employees altered the original file extensions, making format identifications based on their extensions impossible.

There were also several instances where I had to make a best guess as to which file format the extension referred to as some extensions referred to multiple file formats. Once I identified and selected the top 100 formats, I searched for information on relevant version information to include with each format.

After performing the pilot test, I learned that it was too time consuming for participants to rate 100 file formats. After eliciting feedback from the pilot participants on how many file formats they thought they could reasonably rate within the timeframe discussed in the recruitment letter (5-10 hours, in 2-3 rounds, over the course of 4-8 weeks), I decided to reduce the number of file formats to 50. First, I removed from the list file formats that at least half of the pilot participants indicated that they did not know enough about to rate. I was able to remove 45 formats from the list using this criterion. I then removed five additional file formats that half of participants did not know about, starting with the least frequently appearing formats according to the NARA count and moving up. See *Table 3.5.1.1* for a list of the final 50 test formats used in the study.

Extension	Format Name	Version	Instances in NARA Corpus
1. .nc	NetCDF (network Common Data Form)	1.9.1 Classic Format	7,653,015
2. .xml	Extensible Markup Language	1.0	5,710,321
3. .pdf	Portable Document Format	1.0	2,150,735
4. .png	Portable Network Graphic	1.0	1,658,086
5. .kml	Keyhole Markup Language	2.2	1,428,855
6. .txt	Plain Text	No Version Information	1,184,305
7. .csv	Comma Separated Values	No Version Information	1,048,958
8. .gif	Graphical Interchange Format	87a	697,136
9. .html/.htm	Hypertext Markup Language	2.0	571,081
10. .jpg/.jpeg	Joint Photographic Experts Group	Original	421,688
11. .xls	Excel Spreadsheet	5.0	89,191
12. .tif/.tiff	Tagged Image File Format	5.0	62,249
13. .wpd	WordPerfect Document	6.2	29,788

Extension	Format Name	Version	Instances in NARA Corpus
14. .doc	Microsoft Word Document	Word 2000, version 9.0	23,042
15. .sgi	Silicon Graphics Image File	0.97	14,399
16. .hdf	Hierarchical Data Format File	HDF4	14,335
17. .kmz	Google Earth Placemark File	No Version Information	12,329
18. .mp3	Moving Picture Expers Group Audio File	MPEG-2 Audio Layer III	9,056
19. .ppt	PowerPoint Presentation	2.0 for Windows	3,559
20. .mdb	Microsoft Access Database	7.0 for Windows	3,462
21. .spx	Ogg Vorbis Speex File	1.1.12	3,388
22. .mov	Apple QuickTime Move	3.0	3,012
23. .c	C Source Code File	ANSI C	2,310
24. .vsd	Visio Drawing File	6.0	2,305
25. .js	JavaScript File	1.5	2,126
26. .css	Cascading Style Sheet	2	1,836
27. .wk1	Lotus 1-2-3 Worksheet	2.0	1,681
28. xsl	XML Style Sheet	2.0	1,274
29. .raw	Raw Image Data File	ISO 12234-2, TIFF/EP	1,194
30. .rtf	Rich Text Format	1.6	1,180
31. .bmp	Bitmap Image File	5	882
32. .mpg	MPEG Video File	MPEG-1 Part 2	847
33. .docx	Microsoft Word Open XML Document	Word 2007	826
34. .wmv	Windows Media Video File	7	807
35. .wav	WAVE Audio File	Original, no subtypes	588
36. .php	PHP Source Code File - Hypertext Preprocessor	5.0	509
37. .wrl	Microsoft Write	Microsoft Windows 1.0	491

Extension	Format Name	Version	Instances in NARA Corpus
38. .msg	Microsoft Outlook Email Message	Windows Outlook 2007	442
39. .svg	Scalable Vector Graphics	1.1	390
40. .sdw	StarOfficeWriter Text Document	5.0	350
41. .wmf	Windows Metafile	Windows 3.1	327
42. .avi	Audio Video Interleave File	2.0	311
43. .psd	Adobe Photoshop Document	CS	309
44. .for	Fortran Source File	77	302
45. .pptx	PowerPoint Open XML	Microsoft Office 2007	283
46. .swf	Shockwave Flash Movie	5	263
47. .rm	Real Media File	4.01	244
48. .xlsx	Microsoft Excel Open XML Spreadsheet	Microsoft Office 2007	172
49. .pl	Perl Script	5.6	88
50. .ico	Icon File	No Version Information	74

Table 3.5.1.1. The fifty test file formats presented to Delphi participants to rate.

3.5.2. Selecting File Format Endangerment Factors

A review of existing literature was conducted and has revealed many discussions of the importance of assessing a file format's stability for long-term preservation. Several of these discussions include proposed measures for assessing file formats for preservation purposes. Within the literature, I identified a dozen different lists of file format evaluation criteria, displayed in *Table 2.4.1.* I used these criteria lists as the starting point for what eventually became the list of file format endangerment factors rated in the factor-rating Delphi and in the factor-testing questionnaire.

I used a semi-structured method to compile a draft list of factors. I copied each of the evaluation criteria into a document with citations to the original reports for reference. I then compiled all of the factors into one list, removing exact duplicates as I went. This process resulted in a list of nearly fifty factors.

I then started a new list of factors, grouping similar factors together by reviewing provided descriptions. For example, I grouped *widely accepted*, *widespread use*, *popularity*, *market share*, and *adoption* into the factor *ubiquity*. I evaluated each group of similarly themed factors and selected a name for the group that best described them. This process resulted in a list of twenty factors. I then wrote definitions for each of the remaining factors.

I provided a list of all of the factors that were presented in the literature to a knowledgeable friend who independently performed the same task. There were a number of differences in the way this person grouped and named the factors. We met and discussed each of our factor groupings and reached an agreement on the final synthesis of factor lists. The following are the resulting factors and their definitions:

1. **Backward/Forward Compatibility** - whether or not newer versions of the rendering software can render files from older versions, or whether or not older versions of rendering software can render files from newer versions.
2. **Community/3rd Party Support** - the degree to which communities and/or parties beyond the original software producers support the file format.
3. **Complexity** - relates to how much effort has to be put into rendering and understanding the contents of a particular file format.
4. **Compression** - whether or not, and the degree to which a file format supports compression.
5. **Cost** - The cost to maintain access to information encoded in a particular file format, e.g. to migrate files, to maintain the rendering software, or to run an emulation environment.
6. **Developer/Corporate Support** - whether or not the entity that created the original software that produces output in the file format continues to support it.
7. **Ease of Identification** - the ease with which the file format can be identified.

8. **Ease of Validation** - the ease with which the file format can be validated, where validation is the process by which a file is checked for the degree to which it conforms to the format's specifications.
9. **Error-tolerance** - the degree to which this format is able to sustain bit corruption before it becomes unrenderable.
10. **Expertise Available** - the degree to which technological expertise is available to maintain the existence of software that can render files saved in this format.
11. **Legal Restrictions** - the degree to which this file format is or can be restricted by legal strictures such as licensing, copy and intellectual property rights.
12. **Lifetime** - the length of time the file format has existed.
13. **Metadata Support** - whether or not the file format allows for the inclusion of metadata.
14. **Rendering Software Available** - whether or not any type of software is available that can render the information stored in this file format.
15. **Revision Rate** - the rate at which new versions of this file format's originating software are released.
16. **Specifications Available** - whether or not documentation is freely available that can be used to create or adapt software that can render information stored in this file format.
17. **Standardization** - whether or not this file format is recognized as a standard for use and/or preservation by a reputable standards body.
18. **Storage Space** - the average amount storage space a file saved in this format requires when saved.
19. **Technical Dependencies** - the degree to which this file format depends on specific software, operating systems, and hardware in order for its contents to be successfully accessed or rendered.
20. **Technical Protection Mechanism** - whether or not this file format allows for or is encumbered by technical protection mechanisms such as Digital Restrictions Management (DRM).
21. **Ubiquity** - the degree to which use of this file format is widespread and in common use.

3.6. Participants

I conducted a pilot test of the two Delphi questionnaires. Fifteen recruits, colleagues who I knew to have at least some experience and understanding of file formats and digital preservation, agreed to participate in the pilot. Thirteen participants completed the first questionnaire, and twelve participants completed the entire pilot study. I presented participants with the recruitment text and asked for feedback on clarity and wording. All participants indicated that the wording was clear and understandable.

I then invited pilot participants to complete the expertise information questionnaire, where I asked each pilot participant to provided his/her name, email address, occupation, a description of experience with file formats and number of years of experience working with file formats in a digital preservation context. I used the information they provided in this questionnaire to separate them into two groups of six participants, with relatively equal amounts of reported expertise. Furthermore, it became clear that asking for participants' names and email addresses was unnecessary, so I removed those questions from the expertise questionnaire I presented to participants in the actual study.

For the actual study, participants for the two Delphi questionnaires were selected from a group of individuals I identified as having expertise on file formats. Luo and Wildemuth recommended that experts be chosen based on "practical experience in implementing, managing, and evaluating [the desired expertise topic]; research experience in studying [the desired expertise topic]; publications on the topic, and so on" (2009, p. 85). Based on these recommendations, I chose recruits for the Delphi questionnaires who have demonstrated experience in managing and evaluating file formats in a digital preservation environment, conducting research on file formats in digital preservation, and/or producing publications on the topic. These people have demonstrated experience in these areas either through producing publications, giving presentations, teaching workshops or courses, or posting blogs about working with or evaluating file formats in a digital preservation context. Additionally, several people were identified as file format experts by experts already identified for the study. See *Appendix A: Recruitment Text A: for Delphi Participants* for the text I used to recruit participants.

Delbecq, Van de Ven, and Gustafson (1975) recommended that for a homogenous group, ten to fifteen participants is adequate, and it is best if panels do not exceed thirty participants. Accordingly, the aim for this study was to assemble two groups of 10-15 expert participants for the two-phase Delphi portion of the study. I initially recruited a total of 25 participants for the Delphi studies, for a total of 12 participants in one Delphi study and 13 participants in the second Delphi study, which provided a 2.5 participant cushion for attrition.

Of the seventy people I invited to participate in the Delphi studies, a total of twenty-six recruits answered the invitation, indicating that they would like to participate. Twenty-five of the twenty-six responded to the expertise questionnaire. Of these twenty-five participants, four dropped out of the study before the Delphi questionnaire process began. Twenty-one participants completed all or most of the Delphi questionnaires. Twenty participants completed the entire Delphi process. I present here only the data collected from the twenty-one participants who participated in the Delphi questionnaire process. Participants reported a wide variety of job titles, shown in *Table 3.6.1*, which provided some context when examined with their descriptions of their file format experience.

Participants reported file format experience ranging from one to thirty years. The twenty-one participants reported a total of 210 years of working with file formats in a digital preservation context. This results in an average of ten years of experience per participant. The study includes some participants with a comparatively low number of years of relevant experience because of the quality of their reported experience.

Job Titles
Archivist
Information Systems Project Manager
Digital Preservation Analyst
Service manager
Associate director [Digital Library]
Librarian/Programme Director of Preservation Research
Digital Preservation Researcher
Manager Web Archiving Operations
Independent Consultant [on] Recordkeeping
Web Archiving Technical Lead
Research Scientist in Computer Science
Manager of Digital Preservation and Repository Services
Digital Preservation Manager
Director of Research
Digital Preservation Consultant
Electronic Records Format Specialist
Digital Preservation Technical Architect
Associate Professor
Digital Library Services Professional (web archiving)
Library Technical Specialist
Software Engineer / Researcher

Table 3.6.1. Reported participant job titles (references to location removed).

Participants' descriptions of their file format experience typically included specific explanations of how they have worked with file formats. For example one participant wrote, "My main job is to understand digital objects from a structural / technical perspective. As such, I spend most my work time assessing, testing and manipulating file format objects, reading format specs to see how specific implementations of formats stand up." Another participant wrote, "I have researched and evaluated file formats so that I could make recommendations on 'preservation' formats, write preservation plans and make decisions about files that are in need of format migration so that they could remain usable on modern platforms."

In evaluating this text, I made notes on whether participants described their experience as being more dominant in technical analysis of file formats or recommendation and policy generation. For example, the first quote in the paragraph above mentions "testing and manipulating file format objects," which I noted as "technical." The second quote refers to recommending preservation formats and writing preservation plans, which I noted as "recommendation/policy."

I asked the question on the importance of file format endangerment also to answer potential questions of bias in the format-rating questionnaire. The results of this question are displayed in *Table 3.6.2.* I used the data collected from this question, balanced with the quantity and quality of reported expertise, to create a group with a more equal distribution of ratings from this question. The format-rating group had the one participant who rated it as *not important*, three participants who rated it as *somewhat important*, two participants who rated it as *important*, and four participants who rated it as *very important*. The group had a total "importance" mean value of 2.9, which places them between *somewhat important* and *important,* but closest to *important*.

Value	Answer	Response #	%
1	Not important	1	5%
2	Somewhat important	5	24%
3	Important	4	19%
4	Very important	11	52%
	Total	21	100%

Table 3.6.2. Participant ratings of importance of file format endangerment in access to digital information.

I recruited one additional participant to serve as a special reviewer for the fourth questionnaire of the study. This reviewer demonstrated a basic understanding of file formats and the challenges they pose to digital preservation. The reviewer demonstrated an aptitude to be trained for this study and was able to demonstrate skills in searching for information about file formats and for rating file format endangerment levels. The reviewer was trained in a one-on-one session where I reviewed the factors, the file formats, and the data collection guide that I created for him. See *Appendix A: Recruitment Text B: for Special Reviewer* for the text I used to recruit this participant. The 21 people who participated as experts in the two Delphi studies were:

1. Stephen Abrams, University of California Curation Center (UC3) at the California Digital Library (CDL)
2. Micah Altman, MIT Libraries
3. Kevin Ashley, Digital Curation Centre
4. Euan Cochrane, Yale University Library
5. Maurice de Rooij, National Archives of the Netherlands (NANETH)
6. Kevin DeVorsey, National Archives and Records Administration (NARA)
7. Mark Evans, History Associates (formerly at Tessella, Inc.)
8. Carl Fleischauer, Library of Congress
9. Jay Gattuso, National Library of New Zealand
10. Andrea Goethals, Harvard Library
11. Sergiu Gordea, Austrian Institute of Technology
12. Hans Hoffman, Independent Consultant
13. Matt Holden, Institut National de l'Audiovisuel

14. Andrew Jackson, British Library
15. Catherine Jones, Science and Technology Facilities Council
16. Steve Knight, National Library of New Zealand
17. Jerome McDonough, University of Illinois at Urbana-Champaign
18. Peter May, British Library
19. Erin O'Meara, Gates Archives
20. Nicholas Taylor, Stanford University
21. William Underwood, Georgia Tech Research Institute

3.7. Design and Administration of Questionnaire 1

In the first questionnaire I asked all of the recruited participants to provide information related to their experience working with file formats. I asked them to name their occupation, describe their experience working with digital file formats, and list how many years of experience they have had in managing and evaluating file formats in a digital preservation environment and/or conducting research on file formats in digital preservation. Additionally, I asked participants to rate how important they considered file format endangerment to be as a risk factor to the future access of digital materials. See *Appendix B: Questionnaire Designs, Questionnaire 1: Experience with File Formats* to view images of the questionnaire design in the Qualtrics software. Twenty-five participants completed this questionnaire, but four participants dropped out of the study because of illness or other life events before they participated in the Delphi exercise, leaving a total of 21 Delphi participants.

These data allowed me to quantitatively and qualitatively assess the amount of expertise each participant had on file formats and aided me in forming Delphi groups with similar levels of expertise. I placed ten participants in the format-rating group, and eleven participants in the factor-rating group. I used years of experience, “importance” ratings, and the coding from participant experience description text to create the groups. My intent was to balance the three variables (years of experience, importance rating, type of experience) and create two groups with a relatively equal number of years of experience and importance rating.

At the beginning of the questionnaire, participants were presented with the terms of the study that were described in their recruitment letter. They were asked whether or not they agree to these terms, YES or NO. If they answered no, they were taken to an end-of-questionnaire screen that thanked them for their time. If they answered yes, they had given consent to be in the study and they were taken to the beginning of Questionnaire 1.

3.8. Design and Administration of Questionnaires 2 and 3, Round 1

I first conducted a pilot test of Questionnaires 2 and 3. I asked one group of six participants to rate a list of 100 file formats on the first draft of the endangerment scale:

- Information stored in this file format is already inaccessible.
- Information stored in this file format will be inaccessible in 1-5 years.
- Information stored in this file format will be inaccessible in 6-10 years.
- Information stored in this file format will be inaccessible in 11-20 years.
- Information stored in this file format will be inaccessible in more than 20 years.
- I am not familiar enough with this file format to rate it.

I asked participants to write a brief description of the rationale for their answers. I collected each response and their incumbent rationale text into one document and shared the anonymous collected answers with all of participants in the group.

Participants provided feedback on minor spelling errors and inconsistencies with format names and versions. Most importantly, I received overwhelming feedback that rating 100 file formats took far longer than the projected time frame of the study. Based on this feedback and participant input on how long it took them to give thoughtful ratings, I reduced the number of formats to 50. First, I removed from the list file formats that at least half of the pilot participants indicated that they did not know enough about to rate. I was able to remove 45 formats from the list using this criterion. I then removed five additional file formats that half of participants did not know about, starting with the least frequently appearing formats according to the NARA count and moving up.

I asked participants to re-rate the file formats after reviewing the responses of their fellow participants. In this second round of rating the formats in the questionnaire, I reduced the number of formats to 50, based on participant feedback. After participants completed the second round, I computed Spearman's rank correlation coefficient values for each of the file formats and determined that a third round was not necessary.

For Questionnaire 3, I asked the second group of six participants to rate a list of 21 file format endangerment factors on a scale for how relevant the factor is as a cause of file format endangerment:

- Not relevant at all
- Somewhat relevant
- Very relevant

I received very little feedback from pilot participants on the factor-rating questionnaire. Two participants indicated some confusion about being asked to answer the same questions in subsequent rounds of the Delphi. Based on this feedback, I added introductory text to the questionnaires that explained this more explicitly. All other participants indicated that the instructions and purpose were clear, and so I made no other changes to the factor-rating Delphi process based on the pilot.

I compiled and shared the ratings and explanatory text for each questionnaire with the appropriate participants after completing the first questionnaire. I then asked them to review their fellow participants' responses, and then re-take the same questionnaire in a second round. After the second round of questionnaires was complete, I calculated Spearman's rank correlation coefficient values based on the procedures described above. Based on the calculated values, I determined that answer stability was achieved after two rounds, and concluded the pilot at this point.

For the actual study, Questionnaires 2 and 3 were administered to each of their corresponding participant groups using the Qualtrics online survey software. I sent links to the questionnaires to participants using the Qualtrics link generator and email-tracking module. Participants had twelve days to complete the questionnaires and were issued reminders after seven days, ten days, and twelve days.

Questionnaire 2. File Format Endangerment Level Rating. In the second questionnaire, I asked participants to rate a list of 50 file formats according to the degree to which they believed information encoded in each format is at risk of not being accessible. The choices for rating file format endangerment levels for each of the formats consist of a six-point scale:

- Information stored in this file format is already inaccessible.
- Information stored in this file format will be inaccessible in 1-5 years.
- Information stored in this file format will be inaccessible in 6-10 years.
- Information stored in this file format will be inaccessible in 11-20 years.
- Information stored in this file format will be inaccessible in 20 years or more.
- I am not familiar enough with this file format to rate it.

During recruitment, I changed the fifth rating from its original wording of "more than 20 years" to "20 years or more" to more accurately reflect an indefinite time-period after a conversation with one of the recruits. I asked participants to explain the rationale of their ratings for each of the file formats. See *Appendix C: Questionnaire Designs, Questionnaire 2: Rating File Formats,* to view the design of this questionnaire. I provided participants with a short file format-rating guide that provided definitions of key terms used in the questionnaire. The guide appears in *Appendix C: File Format Rating Guide for Participants*.

Questionnaire 3. File Format Endangerment Factor Rating. The third questionnaire was designed to collect expert opinion on which factors are most relevant as causes of file format endangerment. In the questionnaire, I presented participants with the list of file format evaluation factors compiled from the dozen file format evaluation lists found in the literature.

In this questionnaire, I asked participants to rate each factor on an ordinal scale that indicates degrees of relevancy of the factor as a cause of file format endangerment:

- Not relevant at all
- Somewhat relevant
- Very relevant

I also asked participants to provide a brief narrative to explain their ratings for each of the factor options. I also asked participants to provide a brief narrative to explain their ratings for each of the factor options. Additionally, I asked participants to suggest factors that they believed to be a cause of file format endangerment that were not included in the original list, and their rational for suggesting the factors. See *Appendix B: Questionnaire Designs, Questionnaire 3: Factor Rating,* to view the design of this questionnaire.

3.9. Design and Administration of Questionnaires 2 and 3, Round 2

After participants completed their questionnaires, I created documents with participants' anonymous ratings and explanatory narratives for each questionnaire. I shared the documents with participants of the appropriate studies and asked them to review each other's answers and narratives, and to thoughtfully reconsider their original answers. I then asked them to answer a fresh version of their questionnaire in a second round.

Questionnaire 2. File Format Endangerment Level Rating. After the first round of the Format Rating Questionnaire, I removed file formats from the rating list that 50% or more of participants indicated that they did not know enough about to rate. After this process, a total of seven file formats were removed from the list of formats that participants were asked to rate in Round 2 of the Format Rating Questionnaire. See *Table 3.8.1.* for a list of the formats removed from the original list of 50.

Extension	Format Name	Version
.sgi	Silicon Graphics Image File	0.97
.mdb	Microsoft Access Database	7.0 for Windows
.spx	Ogg Vorbis Speex File	1.1.12
.wk1	Lotus 1-2-3 Worksheet	2.0
.wri	Microsoft Write Microsoft Windows	1.0
.sdw	StarOfficeWriter Text Document	5.0
.swf	Shockwave Flash Movie	5

Table 3.9.1. File formats not included in Questionnaire 2, Round 2.

Questionnaire 3. File Format Endangerment Factor Rating. Some participants suggested additional index factors during the first round of the Factor Rating Questionnaire. I reviewed the 16 suggested factors listed below, and selected six new factors that had not in some way been addressed by the original list of 21 factors. I informed participants of the rationale for selecting the final six new factors. For example, one participant suggested, "Existence of a community around the format," however, this factor was already addressed under the factor, *community/3rd party suppor*t. The sixteen factors suggested by participants in Round 1 were:

1. Free specification (not just available specification) - *covered under specifications available and legal restrictions*
2. Support from open source software, - *covered under developer/corporate support and rendering software available*
3. License - *covered under legal restrictions*
4. Native support by common web browsers (e.g. doesn't require plugins) - *covered under technical dependencies*
5. Proprietary - *covered under legal restrictions and specifications available*
6. Existence of a community around the format (especially to ask questions), - *covered under community support*
7. Quality of the specification (how clear is it?) - *sub-factor to specifications available, possible new information, included in new factors*
8. Geographic spread - *possible new information, included in new factors*
9. Domains for specialized formats - *possible new information, included in new factors*
10. The type of the content stored - *possible new information, included under the new factor, Value, which emerged from format rating text*
11. If there are legal regulations affecting a file format - *covered under legal restrictions*
12. If can be rendered on standard (client) PC - *covered under technical dependencies*
13. If can be rendered on a 5 years old PC / (OS) - *covered under technical dependencies*
14. Institutional policies - *possible new information, included in new factors*
15. Vulnerability to viruses - *possible new information, included in new factors*
16. Web vs local access - *possible new information, included in new factors*

Additionally, I evaluated the justification narratives in the first round of the Format Rating Questionnaire for the emergence of additional factors that should be included in the Factor Rating Questionnaire. Based on this evaluation, I added the factor, *value* to the second round of the Factor Rating Questionnaire. I added a total of seven new factors to the second round of the Factor Rating Questionnaire and asked participants to rate them on the same scale as the original twenty-one factors. The following are the seven new factors that I added to the original 21 factors to be rated in Questionnaire 2, Round 2:

1. **Value** - the degree to which information encoded in this format is valued.
2. **Geographic Spread** - the way in which a file format is spread across the world; whether spread thinly across the globe or condensed heavily in a particular area.
3. **Domain Specificity** - the degree to which the format is used only within specific domains.
4. **Viruses** - the degree to which the format is susceptible to containing or being damaged by viruses.
5. **Availability Online** - the degree to which the format is available on the Web.
6. **Institutional Policies** - the degree to which a file format is affected by institutional polices, such as whether or not an institutional policy states that content encoded in this format will be collected and preserved.
7. **Specification Quality** - (sub-factor of "Specifications Available") the understandability and usefulness of the format's available specifications in maintaining access to content encoded in that format.

3.10. Using Spearman’s Rank Correlation Coefficient to Signal Delphi Termination

While typical practice for signaling the termination of the Delphi is for the researcher to use his/her judgment, some research demonstrates the use of a variety of statistics to determine answer stability and/or convergence (Dajani, Sincoff, & Talley, 1979; Kalaian & Kasim, 2012).

I originally used a two-sample chi-square test for independence according to recommendations by Dajani, Sincoff, and Talley (1979) to signal the termination of the Delphi rounds, however, I discovered complications with using this statistic for this purpose. First, the Chi-square statistic assumes that data points from the two distributions are independent; however the two distributions measured in the Delphi study were from the same source, i.e., the same participants. Second, accuracy of Chi-square calculations depends on each cell (number of scores for each item in the Likert-type scale) having at least five observations (Dajani, Sincoff, and Talley, 1979). This second condition was violated for this study due to the low number of participants.

After participants completed the second and third round of Questionnaires 2 and 3, I calculated Spearman's rank correlation of ratings collected between rounds one and two, and between rounds three and four to determine if I needed to administer the questionnaires in additional rounds. I determined answer stability for each question in both questionnaires using Spearman's rank correlation according to recommendations by Kalaian and Kasim (2012). Kalaian and Kasim recommend this statistic for Delphi studies with fewer than 30 participants and with skewed distribution of responses.

After setting the Type I error rate at .05, I calculated the Spearman's rank correlation coefficient using the following formula:

$$r_s = 1 - \frac{n\sum d_i^2}{n(n^2 - 1)}$$

Where:

r_s is the Spearman's rank correlation coefficient,

n is the number of experts,

d, is the difference between the ranks of the ratings on the i^{th} item of the Delphi questionnaire.

I calculated all Spearman's rank correlation coefficients for both the format rating and the factor rating Delphi studies to be between 0.81 and 1 (coefficients for each item are shared in the results section below).

- For the format rating questionnaire, *df* = 10
- For the factor rating questionnaire, original 21 factors *df* = 11
- For the factor rating questionnaire, additional 7 factors *df* = 10

Cross-checking the degrees of freedom value against the Type I error rate of .05 on the Spearman's rank correlation coefficient critical value table, I determined the critical values to be as follows:

- Format-rating critical value = 0.564
- Factor-rating, original 21 factors critical value = 0.536
- Factor-rating, additional 7 factors critical value = 0.564

Since all Spearman's rank correlation coefficient values were above their associated critical values, answer stability was signaled after two rounds of rating for each item in both Delphi studies. Answer stability was achieved after Round 2 for the Format Rating Questionnaire. Answer stability was achieved after Round 2 for the original twenty-one factors in the Factor Rating Questionnaire, and after Round 3 for the seven factors that were introduced in Round 2.

3.11. Design and Administration of Questionnaire 3, Round 3

I asked participants to answer Questionnaire 3 for a third time with only the seven new factors introduced in the second round. This gave participants an opportunity to rate the new factors a second time. As with previous rounds, I collected the anonymized responses into a document and asked participants to review the document as they re-rated the factors. After participants completed this round of the questionnaire, I computed Spearman's rank correlation coefficient values for each of the seven factors and determined that the answers were stable enough to signal the end of the study.

3.12. Design and Administration of Questionnaire 4

The fourth questionnaire was administered to one trained, special reviewer. This questionnaire was designed to address Research Question 4 (How effectively can the expert-chosen file format endangerment factors be applied to rating file format endangerment?) through testing the practicality of applying the factors chosen in the Factor Rating Questionnaire for relevance as causes of file format endangerment. The goal of this final questionnaire was to determine the degree to which a non-expert independent reviewer's responses agree with the output of the Delphi process. In this questionnaire, the reviewer was presented with each of the file formats that were not removed from the Format Rating Questionnaire.

For each file format, I asked the reviewer to:

1. Review a guide on possible data collection sources that I created based on data I collected from the file format rating Delphi questionnaire.
2. Collect and share information from online sources, other recommended sources, or from personal knowledge for each of the factors selected during data analysis of the Factor Rating Questionnaire. (See section *3.5. Data Collection and Analysis: Questionnaire 3*, below)
3. After considering the data collected in step 2, I then asked the reviewer to rate each file format on the file format endangerment level scale used in the Format Rating Questionnaire:
    - Information stored in this file format is already inaccessible.

- Information stored in this file format will be inaccessible in 1-5 years.
- Information stored in this file format will be inaccessible in 6-10 years.
- Information stored in this file format will be inaccessible in 11-20 years.
- Information stored in this file format will be inaccessible in 20 years or more.
- I am not familiar enough with this file format to rate it.

See *Appendix C: Questionnaire Designs, Questionnaire 4: Index Testing,* to view the design of the Factor Testing Questionnaire.

After the reviewer collected factor information for each of the forty-three file formats, I asked him to rate each of the factors using the same scale for relevancy as a cause of file format endangerment that was used in the Factor Rating Questionnaire:

- Not relevant at all
- Somewhat relevant
- Very relevant

Because the special rater had just gone through the exercise of searching for information on each factor and applying this directly to rating the file formats, his ratings were strongly based in the reality of putting the factors to use in a real-world scenario. This activity provided me with additional data that I used to compare with other factor-related data that I collected from the file format rating and factor rating Delphi questionnaires.

3.13. Design and Administration of Special rater Follow-Up Interview

I conducted a semi-structured e-mail interview in which I elicited feedback on the process the special reviewer used to collect information for each factor, how useful he found each factor to be in assessing file format endangerment, and any other thoughts and opinions he had about the process. I asked the reviewer the questions listed below:

Round 1 Questions:

1. How did you go about collecting data for each factor?
2. Were there particular file formats that you had a hard time finding information for? What were they? Where did you look for information on this format?
3. Were there particular factors that were generally difficult to find information for? What were they? Where did you look for information for this factor?
4. Generally, how useful did you find the factors to be in helping you rate the endangerment levels?
5. Were there specific factors that were more useful than others? Less useful?

Follow-up Questions

1. You stated that rendering software available, specifications available, and specification quality were the most useful indicators of endangerment. However, you rated rendering software available, specifications available, ubiquity, and community/3rd party support as very relevant, and specification quality as somewhat relevant. Can you explain the discrepancy between your statement and these ratings?
2. You cited many other factors in your file format rating justification text. Can you explain why you cited the other factors in your ratings rationale, but did not cite them as relevant in your statement or in your factor ratings?

   Examples:

   - **.nc NetCDF (network Common Data Form).** "Good community support of a niche format that is well used in specialized research community. Specification availability as well as its status as a published standard equal a long life.”
   - **.png Portable Network Graphic.** "Ubiquity, open source software, good spec, no legal restrictions.”
   - **.rm Real Media File.** "One company who has clearly lost the streaming battle's early streaming container format. If they go out of business, this could be difficult to render fairly soon without OS emulation and a copy of the player software.”
   - **.wmv Windows Media Video File.** "If DRM, could be inaccessible rather soon, if not, open source renders ought to be able to render it for the foreseeable future."
3. Which sources did you find to be most useful in answering the questions in the questionnaire?

3.14. Data Collection and Analysis

I collected and analyzed data through each of the four questionnaires and through a final interview of the special reviewer at their completion of the Factor Testing Questionnaire. I analyzed the collected data both quantitatively and qualitatively to produce conclusions on a baseline level of file format endangerment and which factors are most relevant as causes/formative indicators of file format endangerment.

Questionnaire 1. Information on Expertise. The recruits who agreed to participate in the study were sent a link to the Expertise Information Questionnaire. I used the data collected from this questionnaire to assess participants' levels of expertise. I used this data to determine in which of the two Delphi questionnaires they would participate, with the goal of creating two groups with homogenous expertise levels. Each group contained a relatively equal number of participants with a relatively equal amount of expertise.

I placed ten participants in the format-rating group, and eleven participants in the factor-rating group. I used years of experience, "importance" ratings, and the coding from participant experience description text to create relevant and relatively equal groups. See the group selection in *Table 3.14.1.* My intent was to balance the three variables (years of experience, importance rating, type of experience) and create two groups with a relatively equal number of years of experience and importance rating.

Group: Format Rating	Years of Experience	Importance Rating
Participant 1	20	4
Participant 2	30	2
Participant 3	3	2

Participant 4	1	2
Participant 5	8	4
Participant 6	12	4
Participant 7	12	3
Participant 8	7	1
Participant 9	5	3
Participant 10	4	4
Total	**102**	**29**
Mean	**10.20**	**2.90**
Group: Factor Rating	**Years of Experience**	**Importance Rating**
Participant 11	20	2
Participant 12	14	2
Participant 13	13	3
Participant 14	2	4
Participant 15	11	4
Participant 16	8	4
Participant 17	13	4
Participant 18	10	4
Participant 19	12	4
Participant 20	3	4
Participant 21	2	3
Total	**108**	**38**
Mean	**9.82**	**3.45**
Grand Total	**210**	**67**
Grand Total Mean	**10**	**3.20**

Table 3.14.1. Participant group assignment.

I addressed the point of relevancy by using participants' experience description text, participant job titles, and publications participants have produced. I categorized the file format rating tasks as requiring more specific technical knowledge, so I selected participants who reported more technical experience in the format-rating group. I categorized the factor rating tasks as requiring more skills and knowledge of evaluating file formats for policy and recommendation development, so I selected participants whose experience description text I noted as "recommendation/policy."

When participants' text revealed experience in both technical understanding of file formats and experience with developing recommendations and policy, I used their job titles and publication history (when applicable) to further assess in which groups they would be best suited. As a final measure, I used the number of years of experience and importance ratings as deciding factors for group placement.

This process was complicated somewhat by participant dropout and recruits who responded to the recruitment letter after the initial deadline. The results, however, reflect the desired balance with only a 0.20-year difference (format-rating group) and 0.18-year difference (factor-rating group) from the grand total years of experience mean, and a 0.30-point (format-rating group) and 0.25-point (factor-rating group) difference in the group mean importance ratings and the grand total mean importance rating.

Questionnaire 2. File Format Endangerment Level Rating. Data for this questionnaire was collected through the iterative rating and feedback process of the Delphi method. I computed Spearman's rank correlation coefficient values on the ratings of each of the file formats after the second Delphi round to test for answer stability, i.e., when participants have statistically ceased changing their answers between rounds. The Spearman's rank correlation coefficient values indicated that the ratings of all file formats had sufficiently stabilized after Round 2, and so I concluded data collection for the File Format Rating Questionnaire at that time.

I calculated a final mean response value for each file format. I also reviewed, coded, and analyzed the justification narratives provided by participants; the results of this are discussed below. I further examined the justification narratives to determine if participants considered certain file format characteristics that are not included in the original list of 21 file format evaluation criteria. One new criterion emerged through the format rating Delphi process, *value*, which I included in the list of file format evaluation criteria to be rated in Round 2 of the Factor Rating Questionnaire. I created a list of forty-three file formats rated in the Factor Testing Questionnaire by eliminating file formats that 50% or more participants indicated they did not know enough about to rate.

Questionnaire 3. File Format Endangerment Factor Rating. Data for this questionnaire was collected through the iterative rating and feedback process using the Delphi method. As with the Format Rating Questionnaire, I calculated Spearman's rank correlation coefficient value for the ratings of each of the original twenty-one factors after the second Delphi round to test for answer stability. The Spearman's rank correlation coefficient values indicated that the ratings of all twenty-one factors had sufficiently stabilized after Round 2, and so I concluded data collection for those factors. I performed the same test for the seven new factors after Round 3, through which I determined that there was sufficient answer stability to conclude the data collection for the Factor Rating Questionnaire.

At the conclusion of this Delphi study, I calculated the final mean response value for each file format evaluation criteria. I ordered each factor based on its mean response value. I used this ordered list to form the list of factors that I asked the special rater to use in the Factor Testing Questionnaire. I used the qualitative data collected from the participant justification narratives to create a guide that I provided to the special rater to use in finding information on the criteria (discussed below). I also coded and analyzed the justification narratives to detect emerging patterns within the text.

Questionnaire 4 and Post-Questionnaire Interview. Factor Testing. One trained, special reviewer completed the fourth questionnaire. For this questionnaire, the participant was required to collect information for the factors selected as a result of data analysis from the Factor Rating Questionnaire, for each of the 43 file formats. Using the guide I created from the Format Rating Questionnaire data, the reviewer searched for information online on each of the resulting thirteen Factor Rating Questionnaire factors. The reviewer then considered the information gathered for each factor when rating each file format on the endangerment scale used in the Format Rating Questionnaire. After the reviewer rated the forty-three formats, I asked the reviewer to then rate each of the factors used in the questionnaire for relevancy as a cause of file format endangerment, using the same scale used in the Factor Rating Questionnaire.

Once the reviewer completed Questionnaire 4, I conducted a semi-structured email interview with the reviewer to collect feedback on the process he used to collect information for each factor, how useful he found each factor to be in assessing file format endangerment, and any other thoughts and opinions he had about the process.

Additionally, I used the data collected from Questionnaire 4 to create a guide for future use. I also reviewed comments collected in the Factor Testing Questionnaire and the post-rating interview data, and used this information in conjunction with the qualitative data collected in the Factor Rating Questionnaire to make final decisions about which factors to include in the dissertation's ultimate product, a proposed file format endangerment index. I used this qualitative data to clarify ambiguous quantitative data and to solidify the rationale for the final factor selection.

Data Comparison. I compared quantitative and qualitative data collected using the four questionnaires in this study. For this part of the data analysis, I made three sets of comparisons. First, I compared the file format ratings collected from the special rater against the mean file format rating means collected from the Delphi participants. Second, I compared ranked file format rating data collected from the Questionnaire 2, Round 2 Delphi study; and ranked file format rating data collected from the special rater using Questionnaire 4. Third, I compared data collected from three sources: 1) factor appearance count from the file format rating Delphi justification text; 2) mean factor ratings from Questionnaire 3, Round 2 Delphi study; and 3) factor ratings from Questionnaire 4.

CHAPTER 4: RESULTS

In this chapter, I present the quantitative results collected through the four questionnaires. I also present analysis of and excerpts from the qualitative information collected through the four questionnaires and the special rater follow-up interview to provide additional context for the findings and conclusions reported.

4.1. Questionnaire 2. File Format Rating

In this section, I present the quantitative and select qualitative data collected through the file format rating questionnaire Delphi process. I present Spearman's rank correlation coefficient values for each file format, together with Round 1 and 2 mean scores and their calculated differences. I have collected and presented mean scores for each file format rated in each of the Delphi rounds, and have presented Round 2 ratings and mean scores for each file format. I present a list of file formats, ranked in order of the highest mean scores (deemed most endangered) to the lowest mean scores (deemed least endangered).

I present description and excerpts of participant justification text for each file format. Additionally I present information about a correlation test between Round 2 mean ratings and Spearman's rank correlation coefficient values. Finally, I present the count and description of mentioned factors in participant justification text.

4.1.1. Delphi Termination Using Spearman's Rank Correlation

After the experts completed the second round of rating file formats, I checked for answer stability for each question using Spearman's Rank Correlation. The calculated correlation coefficient values demonstrated that answer stability was reached for all file format ratings after the second round. These calculations are shown in *Table 4.1.1.1.*

File Formats	R1 Mean	R1 sd	R2 Mean	R2 sd	Mean Diff.	r_s
.nc NetCDF (network Common Data Form)	1.00	0.52	1.00	0.48	0.00	0.99
.xml Extensible Markup Language	1.00	0.32	1.00	0.32	0.00	0.98
.pdf Portable Document Format	1.22	0.57	1.22	0.57	0.00	0.98
.png Portable Network Graphic	1.11	0.47	1.00	0.42	-0.11	0.98
.kml Keyhole Markup Language	1.25	0.67	1.67	1.34	0.42	0.81
.txt Plain Text	1.00	0.00	1.00	0.00	0.00	1.00
.csv Comma Separated Values	1.10	0.32	1.00	0.00	-0.10	0.98
.gif Graphical Interchange Format	1.22	0.57	1.00	0.32	-0.22	0.98
.html Hypertext Markup Language	1.11	0.47	1.10	0.32	-0.01	0.97
.jpg Joint Photographic Experts Group	1.56	1.35	1.63	1.42	0.07	0.99
.xls Excel Spreadsheet	1.89	1.34	1.67	0.85	0.12	0.94
.tif Tagged Image File Format	1.00	0.32	1.00	0.00	0.00	0.99
.wpd WordPerfect Document	1.71	1.14	1.78	0.97	0.07	0.92
.doc Microsoft Word	1.44	0.67	1.56	0.70	0.12	0.99

File Formats	R1 Mean	R1 sd	R2 Mean	R2 sd	Mean Diff.	r_s
Document						
.hdf Hierarchical Data Format File	1.50	0.88	1.57	0.99	0.07	0.98
.kmz Google Earth Placemark File	2.00	1.51	1.75	1.43	-0.25	0.98
.mp3 Moving Picture Expers Group Audio File	1.10	0.32	1.00	0.00	-0.10	0.99
.ppt PowerPoint Presentation	2.50	1.63	2.33	1.29	-0.17	0.89
.mov Apple QuickTime Move	1.43	0.82	1.38	0.74	-0.05	0.99
.c C Source Code File	1.13	0.57	1.13	0.57	0.00	1.00
.vsd Visio Drawing File	1.86	1.16	1.50	0.92	-0.36	0.91
.js JavaScript File	1.29	0.74	1.22	0.57	-0.07	0.98
.css Cascading Style Sheet	1.89	1.34	1.80	1.23	-0.09	0.99
.xsl XML Style Sheet	1.13	0.57	1.20	0.42	0.07	0.97
.raw Raw Image Data File	1.67	1.05	1.57	0.99	-0.10	0.95
.rtf Rich Text Format	1.40	0.70	1.40	0.70	0.00	1.00
.bmp Bitmap Image File	1.11	0.47	1.10	0.32	-0.01	0.99
.mpg MPEG Video File	1.38	0.88	1.22	0.57	-0.16	0.98
.docx Microsoft Word Open XML Document	1.22	0.57	1.22	0.57	0.00	1.00
.wmv Windows Media Video File	1.86	1.16	1.63	1.06	-0.23	0.98
.wav WAVE Audio File	1.00	0.32	1.10	0.32	0.10	0.98
.php PHP Source Code File - Hypertext Preporcessor	1.43	0.82	1.43	0.82	0.00	1.00
.msg Microsoft Outlook Email Message	1.25	0.67	1.11	0.47	-0.14	0.98
.svg Scalable Vector Graphics	1.25	0.67	1.22	0.57	-0.03	0.99
.wmf Windows Metafile	1.50	0.99	1.43	0.82	-0.07	0.95

File Formats	R1 Mean	R1 sd	R2 Mean	R2 sd	Mean Diff.	r_s
.avi Audio Video Interleave File	1.83	1.29	1.71	1.23	-0.12	0.95
.psd Adobe Photoshop Document	1.50	0.92	1.67	1.08	0.17	0.95
.for Fortran Source File	1.17	0.67	1.00	0.48	-0.17	0.98
.pptx PowerPoint Open XML	1.25	0.67	1.11	0.47	-0.14	0.98
.rm Real Media File	1.86	1.34	2.00	1.26	0.14	0.89
.xlsx Microsoft Excel Open XML Spreadsheet	1.22	0.57	1.00	0.32	-0.22	0.98
.pl Perl Script	1.17	0.67	1.15	0.63	-0.02	0.98
.ico Icon File	1.13	0.57	1.13	0.57	0.00	1.00
Overall Mean Difference					**-0.05**	

4.1.1.1. File format rating means, standard deviations, mean differences between rounds, and Spearman's rank correlation coefficient values.

4.1.2. Format Rating Results

In this section I present two tables of data collected during the file format questionnaire Delphi process. *Table 4.1.2.1* shows the distribution of ratings and means for each file format rated in Round 2, ranked in order of the highest mean scores (most endangered) to the lowest mean scores (least endangered). I calculated the mean scores after removing and adjusting for the ratings with which participants indicated they were not familiar enough to rate it.

Rank	Format	Already Inaccessible (5.00)	1-5 years (4.00)	6-10 years (3.00)	11-20 years (2.00)	20+ years (1.00)	Not Familiar (NA)	Mean	sd
1	**.ppt** PowerPoint Presentation	0	1	4	1	3	1	2.33	1.29
2	**.rm** Real Media File	0	0	3	1	3	3	2.00	1.26
3	**.css** Cascading Style Sheet	1	0	0	4	5	0	1.80	1.23
3	**.wpd** WordPerfect Document	0	0	2	3	4	1	1.78	0.97
4	**.kmz** Google Earth Placemark File	1	0	0	2	5	2	1.75	1.43
5	**.avi** Audio Video Interleave File	0	1	0	2	4	3	1.71	1.23
6	**.psd** Adobe Photoshop Document	0	1	0	3	5	1	1.67	1.08
6	**.kml** Keyhole Markup Language	1	0	0	0	8	1	1.67	1.34
6	**.xls** Excel Spreadsheet	0	0	1	4	4	1	1.67	0.85
7	**.wmv** Windows Media Video File	0	0	2	1	5	2	1.63	1.06

Rank	Format	Already Inaccessible (5.00)	1-5 years (4.00)	6-10 years (3.00)	11-20 years (2.00)	20+ years (1.00)	Not Familiar (NA)	Mean	sd
7	**.jpg** Joint Photographic Experts Group	1	0	0	1	6	2	1.63	1.42
7	**.raw** Raw Image Data File	0	0	1	2	4	3	1.57	0.99
8	**.hdf** Hierarchical Data Format File	0	0	1	2	4	3	1.57	0.99
9	**.doc** Microsoft Word Document	0	0	0	5	4	1	1.56	0.70
10	**.vsd** Visio Drawing File	0	0	1	2	5	2	1.50	0.92
11	**.php** PHP Source Code File - Hypertext Preporcessor	0	0	0	3	4	3	1.43	0.82
11	**.wmf** Windows Metafile	0	0	0	3	4	3	1.43	0.82
12	**.rtf** Rich Text Format	0	0	1	2	7	0	1.40	0.70
13	**.mov** Apple QuickTime Move	0	0	0	3	5	2	1.38	0.74
14	**.pdf** Portable Document Format	0	0	0	0	2	7	1.00	0.57

Rank	Format	Already Inaccessible (5.00)	1-5 years (4.00)	6-10 years (3.00)	11-20 years (2.00)	20+ years (1.00)	Not Familiar (NA)	Mean	sd
14	**.js** JavaScript File	0	0	0	2	7	1	1.22	0.57
14	**.mpg** MPEG Video File	0	0	0	2	7	1	1.22	0.57
14	**.docx** Microsoft Word Open XML Document	0	0	0	2	7	1	1.22	0.57
14	**.svg** Scalable Vector Graphics	0	0	0	2	7	1	1.22	0.57
15	**.xsl** XML Style Sheet	0	0	0	2	8	0	1.20	0.42
16	**.pl** Perl Script	0	0	0	1	6	3	1.15	0.63
17	**.ICO** Icon File	0	0	0	1	7	2	1.13	0.57
17	**.c** C Source Code File	0	0	0	1	7	2	1.13	0.57
18	**.msg** Microsoft Outlook Email Message	0	0	0	1	8	1	1.11	0.47
18	**.pptx** PowerPoint Open XML	0	0	0	1	8	1	1.11	0.47
19	**.html** Hypertext Markup Language	0	0	0	1	9	0	1.10	0.32
19	**.bmp**	0	0	0	1	9	0	1.10	0.32

Rank	Format	Already Inaccessible (5.00)	1-5 years (4.00)	6-10 years (3.00)	11-20 years (2.00)	20+ years (1.00)	Not Familiar (NA)	Mean	sd
	Bitmap Image File								
19	**.wav** WAVE Audio File	0	0	0	1	9	0	1.10	0.32
20	**.xlsx** Microsoft Excel Open XML Spreadsheet	0	0	0	0	9	1	1.00	0.32
20	**.for** Fortran Source File	0	0	0	0	7	3	1.00	0.48
20	**.nc** NetCDF (network Common Data Form)	0	0	0	0	7	3	1.00	0.48
20	**.xml** Extensible Markup Language	0	0	0	0	9	1	1.00	0.32
20	**.png** Portable Network Graphic	0	0	0	0	8	2	1.00	0.42
20	**.txt** Plain Text	0	0	0	0	10	0	1.00	0.00
20	**.csv** Comma Separated Values	0	0	0	0	10	0	1.00	0.00
20	**.gif** Graphical Interchange Format	0	0	0	0	9	0	1.00	0.32
20	**.tif** Tagged Image File	0	0	0	0	10	0	1.00	0.00

Rank	Format	Already Inaccessible (5.00)	1-5 years (4.00)	6-10 years (3.00)	11-20 years (2.00)	20+ years (1.00)	Not Familiar (NA)	Mean	sd
	Format								
20	**.mp3** Moving Picture Expers Group Audio File	0	0	0	0	10	0	1.00	0.00

Table 4.1.2.1. Round 2 file format rating distribution and mean scores, arranged by mean rank

As is evident in *Table 4.1.2.1*, the distribution of ratings across formats indicates a fair amount of variation, but overall, most of the answers gravitated toward the less endangered end of the scale. Anomalous ratings appear for the formats, .kml, .jpg, .kmz, and .css which have single ratings of “already inaccessible” where the majority of the ratings appear lower in the scale.

Table 4.1.2.1 also shows each of the file formats ranked by their mean scores from Round 2 ratings. Higher means indicate a higher endangerment level rating. The highest rating, 2.33 (.ppt) indicates an endangerment level rating between “11-20 years” and “6-10 years” of when participants estimated content encoded in the format will be inaccessible. Only the .ppt format and .rm formats were rated at or above 2.00. Ten formats (23%) had mean ratings of 1.00, the lowest rating level. The remaining 31 formats (72%) had mean ratings between 1.00 and 2.00.

4.1.3. Format Rating Justification Text

This section contains the results of qualitative analysis of the justification text provided by participants during the Delphi rating process. Most of the text represented here is from Round 2 of the Delphi, and was selected from Round 1 answers in cases when participants referenced their Round 1 justifications in their Round 2 explanation. References to their Round 1 responses were particularly common when participants indicated that they had not changed their ratings in the second round. Within the quoted sections of text I corrected obvious typos and misspellings to preserve the continuity of the text.

Participants included some kind of justification text with their answers. When participants selected the choice, “I am not familiar enough with this format to rate it,” in 82 of 142 (58%) such responses across both rounds, they made short statements like, “not familiar” or “not familiar enough with this format to rate it” to reiterate their selection. The remaining 60 (42%) included more substantive explanations for their answers such as, “I've never poked at HDF and am not comfortable guessing at how long information will remain accessible when encoded in it,” and, “Further information is required in terms of documenting old formats from Microsoft. Responses seem to vary on how well it’s supported. Not enough information to comment on its long term accessibility.”

As part of the questionnaire design, participants were required to provide a text response for each scale response, so there were at least brief text responses for each file format in both rounds. All except one participant included in-depth explanations for their ratings in Round 1. The one participant who did not include in-depth explanations supplied one, repeated answer for every format, except for the .kmz and .kml formats: “We have software that can interact with the format and can maintain access to that software indefinitely.” Participants also included in-depth text responses in Round 2 except in the cases when they referred to their Round 1 explanations. I coded and analyzed Round 1 justification text to detect and count which factors participants discussed in their rationales for their format ratings.

1. .nc. NetCDF (network Common Data Form), Version 1.9.1

Three participants indicated that they were not familiar enough with the format to rate it. All seven of the other participants indicated that information stored in this file format will be accessible for 20 years or more. The reasons they provided for this rating is that the file format has open specifications (3 participants), available software that can render the format (3 participants), and extensive support from the Unidata community (4 participants), a collective, supported by the National Science Foundation, of 250 earth-system education and research organizations.

2. .xml, Extensible Markup Language, Version 1.0

Only one participants indicated that he/she was not familiar enough with the format to rate it. All of the nine other participants rated the format as information being stored in it will be accessible for 20 years or more. The main reasons for selecting this rating are that the XML format is text-based, ubiquitous, has open specifications, is standardized by the W3C, and can be rendered by a large number of software applications. Although 90% of participants indicated that information encoded in XML will be accessible in 20 years or more, six participants noted specifically that while the text of an XML file will be readable, it is unclear whether the meaning of XML files will be retained. One participant stated, “A greater risk is loss of the schema that validate semantic structure for use, with less widely-used schema being at greater risk.”

3. .pdf, Portable Document Format, Version 1.0

70% of participants rated this file format as being accessible for 20 years or more. Four participants mentioned the fact that the PDF version 1.0 is much simpler than later versions and is therefore less difficult than later versions to maintain access to. One participant stated, “… version 1.0 was relatively simple and is well documented and remains accessible through current applications (I think).” Overall, the .pdf format is a complex preservation format, as illustrated by another participant statement that

> “Answering the accessibility question for PDF is more complicated because the format allows for the embedding of arbitrary binary data. So, at the worst, the accessibility of the contents of a PDF could be as poor as the most inaccessible format…”

4. .png, Portable Network Graphic, Version 1.0

The PNG format was also considered to be low risk by participants. The eight participants who knew enough about the format to rate it considered information stored in .png format to be accessible for 20 years or more. Reasons they cited for this rating were available rendering software (4 participants), good documentation (5 participants), popularity (5 participants), and the fact that it is a W3C standard (2 participants). One participant stated that, “As is noted, PNG is a well documented, open format with wide application support. It is also relatively simple when compared to other raster formats so even if all applications disappeared a good programmer could most likely write something from scratch that would be able to interpret PNG files."

5. .kml, Keyhole Markup Language, Version 2.2

The KML format is somewhat anomalous in that everyone who rated the format put it at low-risk, except for one participant, who indicated that information stored in it was already inaccessible. This participant stated, "This is a tough one. It depends on what you consider the role of kml files to be. The objects that they are used to capture information about include web-based components (e.g. map layers) that may already have changed and/or been lost. Therefore the objects themselves no longer exist, just the kml file components of the objects."

The other eight participants indicated that the file format will be accessible for 20 years or more. The prevalent reasoning is that data stored in .kml is an XML schema, is text-based, and is easily parsed into a human-readable form. Additionally, participants indicated that the format was popular, well documented, and has been adopted as a standard by the Open Geospatial Consortium. As one participant wrote, "This is an example of a popular xml schema that is likely to persist, on account of documentation and wide adoption."

6. .txt, Plain Text, No Version Information

According to participants, plain text files are highly ubiquitous, very simple, well documented, and with a wide range of software that can render them. Consequently, all ten participants rated it as being inaccessible in 20 or more years. One participant stated that, "As everyone noted, plain text is pretty much the most ubiquitous format there is and that while different character encodings might be used, this fact should not impede access to the information stored in files of this type."

7. .csv, Comma Separated Values, No Version Information

As with the .txt format, all participants rated the CSV format to be safe from inaccessibility beyond 20 years. Seven participants rated this format in this way, citing the fact that .csv files are text-based files that use a comma to separate individual pieces of data. Additionally, participants indicated that the format was ubiquitous (2 participants), simple (2 participants), standardized (3 participants), and well documented (2 participants).

8. .gif, Graphical Interchange Format, Version 87a

While one participant said he/she did not have enough experience to rate this format, the nine other participants rated it as the lowest risk choice, "20 years or more." As justification for these ratings, participants indicated that it can be rendered using a wide variety of software applications (4 participants), it is very popular/ubiquitous (5 participants), has available specifications, despite its proprietary nature (2 participants).

9. .html, Hypertext Markup Language, Version 2.0

In contrast to all of the formats discussed previously, the HTML format had one rating that indicated that information stored in it would be inaccessible in 11-20 years. However, the justification text that accompanied this rating appears inconsistent with the numeric response: “As with XML, the encoding is basically human interpretable text defined by tags. The tags are defined by external schemas so as long as they are maintained and remain accessible, HTML should be interpretable for at least 20 years.” Considering the fact that this participant stated that he believed the format will be “interpretable for at least 20 years” and the fact that the nine other participants rated it as being accessible in 20 years or more, the overall participant consensus is that the .html v. 2.0 format will be accessible for 20 years or more.

However, this is also a format that is accessible in simple text format and could be renderable in its pure text form over a long time scale, but the proper interpretation and “performance” of all of the elements included in a file is dependent on the web browser that is used to render the file. For example, one participant wrote, “As a general format, HTML is almost entirely backward compatible, and even if it falls from use, the sheer volume of valuable information stored in this format means it will keep being used. Note, however, that specific features (like the blink tag) are already 'obsolete', in that they require additional knowledge and software to know that they are there and to render them.” Another participant similarly wrote,

> HTML 2.0 is already an obsolete standard. However, it's a very simple imperfect subset of a very-widely-used current standard. Rendering of some HTML 2.0 documents is already potentially slightly flaky, but adequate results are still likely to emerge. In any event, I don't think one needs to be able to 'render' HTML (as a browser would) for its simpler forms to be accessible. One can treat it as text and still have an acceptable result.

10. .jpg, Joint Photographic Experts Group File, Original Version

The ratings for this format are somewhat of an anomaly in that participants had different interpretations of which format they were rating. One participant rated the format as already being inaccessible for the reason that he/she was rating the JPEG Interchange Format (JIF) because the survey specified the “original” version. This participant wrote, “Assuming you mean JPEG Interchange format (JIF) as specified in ITU-T Recommendation T.81: Information Technology -- Digital Compression and coding of continuous Tone Still Images - Requirements and guidelines, Sept 18, 1992. This specification is the same as ISO/IEC 10918-1:1994. Open Standard. However, The JIF file format was never widely used. The compression method was however implemented in file format DNG 1.1.”

Eight of ten participants rated later versions of the format that follow the 1994 ISO standard and later. One participant defended his rating of 20+ years by writing, “A couple of respondents noted that jpeg is generally a compression method, not a format but regardless, it is one of the most common formats in existence (pretty much every cell phone creates and renders them) and so should remain accessible for at least 20 years.”

11. .xls, Excel Spreadsheet, Version 5.0

The Microsoft Excel file format also received a diverse range of ratings. Unlike all of the previous formats, there is no single rating that has more votes than the others. In the first round, three participants rated the format as being inaccessible in 1-5 years, but in the second round, there were no ratings for this time period. One of these participants upgraded his/her rating to 6-10 years and explained this change in ratings by writing, “Based on the other comments I increased the amount of time that I believe this format will be accessible. A lot depends on what functionality is present in individual files and how well applications cope with them. As one person pointed out, there is a free viewer that could be preserved moving forward so at least a read capability should be possible."

However, one participant who rated the format at 11-20 years, indicated that, "I'm downgrading my optimism about this format slightly after reading round 1 respondents' comments about the requirement that complex software with many dependencies be preserved in order for this format to remain accessible and the introduction of VBA scripting in version 5.0."

This format also brought up questions about the meaning of the term "accessibility." Does accessibility mean that one can open a file and see the contents in a simplified form; or does it mean that all of the components, including embedded scripts, are functioning as originally intended? One participant indicated that software like the free Microsoft Excel Viewer allows for continued, full access to these files. Another argued that the concept of emulators negates the purpose of this questionnaire: "I don't disagree that virtual machines will allow some form of access in 20 years time - but this is true of any format, and that would make this questionnaire pointless."

12. .tif, Tagged Image File Format, Version 5.0

The TIFF file format proved to be much less complicated than Excel. All ten participants rated it as being accessible for 20 years or more. Reasons participants supplied for this rating include that the format is widely adopted (5 participants), well documented (5 participants), and many software applications are available that can render it (4 participants).

13. .wpd, WordPerfect Document, Version 6.2

Like the Microsoft Excel format, the WordPerfect 6.2 format also had highly variable ratings, though the 20+ years rating had just one more vote than the others. The two participants who rated the format at 6-10 years cited its proprietary nature, shrinking market share, and sharp decrease in use as their rationales. The four participants who rated it for 20+ years cited the existence of open source software that can currently render the format as their reason for this rating. The three participants who rated it at 11-20 years indicated uncertainty over how long the format would receive support from open source rendering projects. With a mean rating of 1.78, it was rated with the third highest endangerment level of the forty-three formats that were rated in the second round.

14. .doc, Microsoft Word Document 2000, Version 9.0

With a mean rating of 1.56, Microsoft Word, version 9.0 is ranked the 11th most endangered file format on the list. This mean rating places .doc just over halfway between 11-20 years and 20+ years. Five participants rated it at 11-20 years. These participants expressed concern about how well and long Microsoft will support this format as well as some of the complexities inherent to the format. One participant summarized the situation well:

> Although this version of the Word format has been effectively reverse-engineered it has never been clearly and formally documented, and to some extent the code is the only real documentation. The other applications which can deal with the format (openoffice, libre office et al) are themselves large and complex software suites with many dependencies on other software environments that are unlikely to survive. MS [Microsoft] themselves are not motivated to provide backwards compatibility for more than one or two versions.

The four participants who rated it at 20+ years provided rationale for their ratings by citing its ubiquity and the availability of open source software and emulators that can be used to render it.

15. .hdf, Hierarchical Data Format File, Version 4

There was a high level of variation in the ratings for the Hierarchical Data Format. Originally, there were three respondents who indicated they did not know enough to rate it, three who rated it at 11-20 years, and four at 20+ years. One participant downgraded his/her rating to 6-10 years. This participant wrote that he/she changed his/her answer based on other participants' justifications, but did not highlight what reasons cause the rating change. One rationale provided in Round 1 that indicates risk was:

> It's difficult to be certain about this, but HDF4 is already a poorly-supported format partly because of poor design decisions in the format itself. HDF5 is designed to address these and the user community that exists is strongly motivated to abandon HDF4 for this reason. There will then be little reason to maintain support for HDF4 in those applications which exist, although I suspect a converter will continue to be available, though conversion is imperfect.

The four participants who rated it at 20+ years cited a large user base ("It has several million users"), that it is an open format, and a strong user community ("The HDF Group supports tools for writing and accessing the hdf4 files (www.hdfgroup.org/)") as rationale for this rating.

16. .kmz, Google Earth Placemark File, No Version Information

The ratings and information collected about the KMZ file proved to be useful. The mean rating value was 1.75, meaning that it was rated between the 20+ year and 11-20 year ratings, but closer to 11-20 years rating. As with the .kml file, one participant rated it to be already inaccessible for the same reason: "It depends on what you consider the role of kmz files to be. The objects that they are used to capture information about often include web-based components (e.g. map layers) that may already have changed and/or been lost. Therefore the objects themselves no longer exist, just the kmz file components of the objects."

Of the eight participants who knew enough about the format to rate it, five indicated that it would be accessible for 20+ years. These participants cited that accessibility was dependent on support for Zip 2.0 in order to open the files. The two participants who rated the file format at 11-20 years indicated that accessibility was also dependent on the .kml files within the container. (The .kmz format is a compressed container format that can contain one or more .kml files.) The .kmz format was ranked 4th most endangered with a mean of 1.75.

17. .mp3, Moving Picture Experts Group Audio File, MPEG-2 Audio Layer III

In the second round, all ten participants rated it as being inaccessible in 20 years or more. One person rated it at 11-20 years in the first round, but changed to 20+ years in Round 2. The participant's rationale for the 11-20 year rating in Round 1 was, "MP3 is a format for encoding audio that supports at least distinct codecs and a number of different tagging methods. While the format is well documented, it is conceivable that certain encodings could prove problematic and require earlier intervention."

Rationale for the 20+ year ratings included a high level of adoption (6 participants), a wide variety of rendering software (3 participants), available documentation (4 participants), and ISO standardization (2 participants).

18. .ppt, PowerPoint Presentation, Version 2.0 for Windows

The PowerPoint 2.0 file format was rated to be the most endangered format, with a mean of 2.33 (between 11-20 years and 6-10 years). This format's ratings also exhibit a high level of variation. 6-10 years received the highest number of ratings (4 participants), and justifications for this rating included no clear existence of specifications (1 participant) and a reliance on Microsoft to provide continued support for this version of PowerPoint (3 participants). Those who rated the format's inaccessibility at 1-5 (1 participant) and 11-20 (1 participant) years also indicated its proprietary nature and dependency on Microsoft's continued support for long-term accessibility as factors contributing to their rating choices. The three participants who rated it at 20+ years cited its ubiquity, the existence of rendering software, and the ability to maintain rendering platforms over time as their rationales for this rating.

19. .mov, Apple QuickTime Movie, Version 3.0

The QuickTime Movie file format was rated with a mean of 1.38, slightly more than the 20+ rating. While five participants rated it as 20+, three of them rated it at 11-20 years. The three who rated it as this level justified their rating by citing a tendency of Apple to drop support for older formats. One participant cited a “gut feeling”: “Although already a legacy format support for this is widespread in many current rendering systems. This is only likely to be dropped if it presents a problem to port to newer environments at a time when there is less legacy content in the format. 11-20 years is a gut feeling for when this might happen.” Participants who chose 20+ years cited ubiquity (1 participant) and the availability of rendering software (3 participants) their reasons for this rating.

20. .c, C Source Code File, ANSI C

The C Source Code format was the first of several source code formats that presented a particular type of challenge to participants. I included source files in the list of test formats because they conformed to the definition of file formats on which I based this research: the “internal structure and/or encoding of a file which allows it to be interpreted or rendered in human accessible form.” Two participants rated this format based on the ability of a text editor to read the contents of the file.

Four participants discussed the availability of compilers for .c files in their rating justifications. One wrote, “I choose to interpret it as meaning that code can be compiled and executed in a contemporary computing environment. … In addition, I still have no trouble building and running FORTRAN 77 programs 36 years on. C is unlikely to be different.”

21. .vsd, Visio Drawing File, Version 6.0

Five participants rated this format at 20+ years. Participants cited the existence of reverse-engineered renderers (3 participants), and a general technical capability to maintain access to content stored in this format over time (3 participants) as their rationales for this rating. The two people who rated it at 6-10 and 11-20 years cited its proprietary nature and unknown intention of Microsoft to continue to provide support as their rationale for these ratings.

22. .js, JavaScript File, Version 1.5

This format had a mean rating of 1.22. Seven participants rated this format at 20+ years. Similar to the C source code files, participants cited the fact that .js files are written in plain text and the text itself should be accessible indefinitely (3 participants). Some cited JavaScript's ubiquity (3 participants) and the availability of JavaScript engines to interpret the format (2 participants) as rationale for this rating. Those who rated it at 11-20 years indicated that there might be difficulty in sustaining the rendering of files' intended behaviors across browsers and over time, which caused them to choose a higher endangerment level while rating the format.

23. .css, Cascading Style Sheet, Version 2

The Cascading Style Sheet file format was rated the second highest on the endangerment scale with a mean rating value of 1.80. While half of participants rated it at 20+ years, the four 11-20 year ratings, combined with the one rating that the content stored in this format is already inaccessible, pushed the mean rating closer to 2.00. It is important to note, however, that even while CSS has the second highest mean rating for endangerment, it still is not over 2.00, resulting in a mean ranking just below the 11-20 years rating.

The participant who rated the format as already inaccessible, wrote, “CSS Version 2 RFC is no longer maintained by W3C. Not all browsers correctly parse CSS version 2 code. CSS 2.1 fixes errors in CSS 2 and removes poorly supported features. CSS 2.1 became a W3C recommendation in 2011.” The four participants who rated the format at 11-20 years cited that while CSS files are written in simple text, which they had rated as being available past 20 years, they believe that browser incompatibility (3 participants) and potential disassociation of the file with its referenced web page(s) (2 participants) can prevent meaningful access to the contents in the file.

24. .xsl, XML Style Sheet, Version 2.0

Eight of ten participants rated the XSL format at 20+ years. Participants indicated that XSL files, “contain XSLT templates for transforming, typically, XML based documents into other XML.” Four of these participants indicated that since the contents of the format are written in simple text, it should be accessible in the long-term. Two participants cite that it is a W3C open standard and there are a number of open source applications that can sufficiently render its contents. The two participants who rated the format at 11-20 years cited that version 2.0 is not widely used and the possibility that an XSL file being separated from its referenced XML may reduce its accessibility over time.

25. .raw, Raw Image Data File, ISO 12234-2, TIFF/EP

Four participants rated this format at 20+ years, two at 11-20, one at 6-10, and three did not know enough about the file format to rate it. This format ranked as the 8th highest endangerment level with a mean rating value of 1.57. Those who rated it at 20+ years cited availability of specifications (4 participants), and that Raw Image Data files are a sub-type of the TIFF format, which they also rated at 20+ years (3 participants). Of the two participants who rated it at 11-20 years, one cited difficulty in finding rendering applications and another cited possible use of proprietary compression methods as rationale for a higher endangerment rating. The participant who rated it at 6-10 years indicated that the format has, “Limited use compared to other alternatives such as DNG or EXIF/DCF.”

26. .rtf, Rich Text Format, Version 1.6

Seven of the ten of participants rated it at 20+ years, though two rated it at 11-20 and one at 1-6, which accounts for its 1.4 mean rating value. Those who rated it at 20+ years cited available specifications (3 participants), the availability of software that can render it (3 participants), and ubiquity (2 participants) as reasons for their ratings. One participant who rated it at 11-20 years indicated that while the specifications are available for this format, they are not complete. The participant who rated it at 6-10 years cited concern about Microsoft’s continued support of the format and the fact that it is not as widely used as .doc and .docx formats.

27. .bmp, Bitmap Image File, Version 5

The Bitmap Image file is another format with straightforward ratings. Nine of the ten participants rated it at 20+ years. Rationale for these ratings includes the availability of documentation (4 participants), simplicity (5 participants), and availability of software to render it (6 participants). The participant who rated it at 11-20 years wrote, “Relatively simple format structure, but not promulgated by an independent standards body.”

28. .mpg, MPEG Video File, MPEG-1 Part 2

Seven of the nine participants who rated the format rated it at 20+ years, two rated it at 11-20 years, and did not know enough about the format to rate it, for a total mean value of 1.22. Those who rated it at 20+ years cited an available documentation (5 participants), the availability of rendering software (4 participants), and standardization (3 participants) as their rationale. Of the two who rated it at 11-20 years, on cited the complexity of video formats in general, and the other cited the community’s lack of experience with video formats as reasons. As one participant stated, “Despite the very wide availability of renderers for this format, and its standardization, we have less experience to go on with digital video formats than with digital image formats. I'm therefore being somewhat more cautious on how long this will be easy to deal with, but on very little real evidence.”

29. .docx, Microsoft Word Office Open XML Document, 2007

The .Microsoft Word Office Open XML format received the same distribution of ratings as the .mpg format, again resulting in a 1.22 mean value. The seven 20+ year ratings were justified by citing its wide adoption (four participants), its available documentation (2 participants), its standardization (3 participants), the availability of rendering software (4 participants), and its support from Microsoft (2 participants). The two participants who rated it at 11-20 years cited its complexity as their primary reason for a higher endangerment level rating. One participant stated, “Word Open XML is an xml based word processing format that has been published as an ISO spec and that is well supported. That said, it is complex and so it would be smart for collecting institutions to monitor changes in support and decide on when to migrate information to a newer format.”

30. .wmv, Windows Media Video File, Version 7

The Windows Media Video File format had a much greater variation in ratings and a relative higher mean than the other formats. With a 1.63 mean rating value, it co-ranks with the .jpg format as the 7th highest endangerment level rating. Five of eight participants who rated the format selected the 20+ years timeframe. They justified their answers by citing ubiquity (2 participants), available documentation (2 participants), and available software that can be used to render it (4 participants). The three participants who rated the format with shorter timeframes (11-20 and 6-10 years), indicated that factors such as its proprietary nature (2 participants), a general lack of experience with preserving access to video formats (1 participant), and potential lack of support for this version as newer versions are released (2 participants) were reasons for their ratings.

31. .wav, WAVE Audio File, Original, no subtypes

Nine of ten of participants rated the WAVE Audio File format at 20+ years. These participants cited ubiquity (6 participants), available documentation (5 participants), available rendering software (4 participants), and freedom from licensing encumbrances (4 participants) as their reasons for rating it as such. The one participant who rated it at 11-20 years provided many reasons why the format would remain accessible, but did not provide clear justification as to why he chose the shorter timeframe of availability. One participant did note that it was technically a proprietary format, but suggested that free availability of specifications could mitigate this potential liability.

32. .php, PHP Source Code File - Hypertext Preprocessor, Version 5.0

As with the other source code formats listed here, the PHP format posed some challenges to participants in terms of what "accessibility" meant. In their comments, five participants discussed the differences of rating it as a text file and rating it as whether or not the script can be run. One participant wrote, "Server side scripting language. Source code is plain text, and so accessible in text editor. Execution is interpretation through a PHP engine, which may make running the code difficult in the future. If just interested in being able to view the text, then this format should be accessible for 20+ years; if execution is to be considered, then accessibility may be reduced." Based on the comments, all four of the participants who rated it at 20+ years based their ratings on considering the format as a simple text file. For example, one participant wrote, "I've interpreted 'information stored' as 'being able to get at the contents of the file', which will be straightforward. Actually *running* the PHP may be more difficult."

33. .msg, Microsoft, Outlook Email Message, Windows Outlook 2007

Eight of ten participants rated this format as being accessible for 20+ years, one at 11-20 years, and one did know enough about the format to rate it. The eight participants who rated it at 20+ years cited ubiquity (three participants), available documentation (3 participants), simplicity (2 participants), and availability of rendering software as justification for these ratings. The participant who rated it at 11-20 years noted that though it is ubiquitous, the fact that is a proprietary format may inhibit access in the future.

34. .svg, Scalable Vector Graphics, Version 1.1

Seven of nine participants rated the Scalable Vector Graphics format at 20+ years, and two at 11-20 years, resulting in a 1.22 mean rating value. Those who rated it at 20+ years indicated that their rationale was supported by available documentation (3 participants), its standing as a W3C standard (2 participants), and the availability of rendering software (4 participants). One participant who rated the format at 11-20 years characterized SVG as “inheriting the beneficial properties of XML and standardized through W3C, but not as widely supported as other schemas.”

35. .wmf, Windows Metafile, Windows 3.1

Three participants indicated that they were not familiar enough with the WMF format to rate it. The remaining seven participants were nearly divided between 20+ years (four) and 11-20 years (three). Several participants compared it to SVG in that it is a container format that can contain both vector and raster graphics.

One participant who rated it at 20+ years wrote, “Windows Metafiles may contain vector graphics and raster graphics and thus are similar to SVG files. … They are viewable with the current version of Quickview Plus.”

One participant who rated the format at 11-20 years wrote,

> The dependence of WMF on the functionality of system calls in Windows makes it the only one of the image formats I've been asked to rate that gives me some cause for concern. Although open specifications exist, the specification for this version of the format has been shown to be flawed in some respects. I think many files will still be renderable or convertable beyond 20 years but some may show up the flaws in the spec, and access via an emulation environment with Windows 3.1 may be the only option.

36. .AVI Audio Video Interleave File, Version 2.0

There was a high level of variation in the responses for the AVI format. Three participants indicated that they did not know enough about the format to rate it. One participant rated it at 1-5 years, writing,

> The AVI format is a complex container and allows for such a huge range of codecs to be used within it that one cannot with confidence say that all AVI files will remain accessible for a given period of time. The answer I have given is a worst case - some, probably unusual, AVI files will become unusable in a relatively short timeframe but many are likely to be accessible over much longer timescales, up to 20 years or perhaps more.

One of the two participants who rated the format at 11-20 years cited similar reasoning as the participant who rated it at 1-5 years: “Container format, enabling a range of codecs to be used. Access would depend on availability of appropriate codec, so rather arbitrary timescale selected.” Of the four participants who rated it at 20+ years, some cited as their reasons that it is common (3 participants), and that it has specifications available (2 participants).

37. .psd, Adobe Photoshop Document, Creative Suite (CS)

As with AVI, the Adobe Photoshop there was a large amount of variation among the ratings. Of the eight who rated the format, five rated it at 20+ years. Participants cited its ubiquity (3 participants), available documentation (2 participants), and available rendering software (3 participants) as factors that will maintain its accessibility over time. Of the three who rated it at 11-20 years, two cited its dependency on Adobe for continued support and access. The participant who rated it at 1-5 years wrote, "PSD is proprietary but well documented. I'm not sure what kind of third party support exists though so consider it at a higher risk than other formats." With a mean rating value of 1.67, it shared the ranking of 6th highest endangerment level with two other formats: KML and Excel.

38. .for, Fortran Source File, Version 77

In comparison with other programming language files rated in this study, there was little ambiguity in the participant ratings of the Fortran Source File format. While three participants did not know enough about the format to rate it, all seven participants who did rate it, rated it at 20+ years. One participant noted that as a simple text file, content stored in Fortran files would be accessible in the long-term. The remaining six participants noted that beyond Fortran files being encoded as simple text, they would be able to be compiled and run in the long term as well. One participant justified this by writing,

> FORTRAN is a widely used scientific programming language. International Standard (http://www.fortran.com/F77_std/rjcnf.html) Fortran 77 programs will compile in Fortran 90 compilers. Fortran 77 programs with minor changes will compile in current Fortran compliers.

39. .pptx, PowerPoint Office Open XML, Microsoft Office 2007

Unlike its predecessor, the.ppt format (which had the highest mean endangerment rating at 2.33), the PowerPoint Office Open XML format's mean rating was 1.11. Eight participants rated it at 20+ years citing ubiquity (4 participants), the availability of rendering software (3 participants), availability of documentation (1 participant), and its status as an ISO/IEC standard (2 participants) as rationales for their rating. The participant who rated the format at 11-20 years cited available specifications, available rendering software, and standardization as reasons it would remain accessible, but did not provide rationale for rating it at a shorter timescale.

40. .rm, Real Media File, Version 4.01

The Real Media file format had the second highest mean endangerment rating at 2.00. This places it solidly at the 11-20 years rating. The only other format that had a higher mean rating was PowerPoint at 2.33. There was also a high level of variation in ratings indicating a high level of disagreement in the ratings. Of the three who rated it at 20+ years, one participant indicated that it was a borderline case, though software is currently available that can render it. The three participants who rated it at 6-10 years cited its proprietary nature (1 participant), a shrinking user base (2 participants), lack of clear documentation (1 participant), and an "uncertain commercial future for Real Networks" (1 participant) as rationales for this rating. Two participants commented that there was conflicting information on the Web about which versions of Real Media were documented which made it difficult for them to rate this format. One of these participants rated it at 6-10 years and one at 11-20 years.

41. .xlsx, Microsoft Excel Open XML Spreadsheet, Microsoft Office 2007

Like its companion formats, .docx and .pptx, the Microsoft Excel Office Open XML format received an unambiguously low mean rating of 1.00. All nine participants who rated the format rated it at 20+ years. Their rationale for this rating included its ubiquity (3 participants), the availability rendering software (2 participants), standardization (3 participants), and the availability of specifications (4 participants). One participant summed up the rationale by writing, “As with .docx, I think there is a large enough amount of content available for there to be sufficient pressure to maintain some ability to read this, and the availability of an admittedly complex standard should be sufficient to allow this to happen.”

42. .pl, Perl Script, Version 5.6

Six of the seven participants who rated this format rated it at 20+ years, resulting in a mean value of 1.15. Three participants who rated this format in this way cited the fact that Perl Script is saved in a simple text format and will be accessible indefinitely. One participant did explain that the language was well documented which would allow for understandability in the future. The one participant who rated it at 11-20 years cited ubiquity and available documentation as reasons the format will be accessible, but did not provide rationale for the shorter timeframe rating.

43. .ico, Icon File, No Version Information

Seven of the nine participants who rated this format, rated it at 20+ years. Those who rated it as such cited ubiquity (3 participants), and availability of software for rendering (5 participants). The one participant who rated it at 11-20 years did not provide rationale for the shorter timeframe rating.

4.1.4. Removed File Formats

In the first round of rating file formats, there were seven file formats that half or more of participants indicated that they did not know enough about to rate, shown in *Table 4.1.4.1*. There was not sufficient information to justify having participants rate them again in Round 2. Though I removed these seven formats from the second round of the Delphi process, the information collected in the first round provides some useful insights. I included the data collected in Round 1 below and discuss implications of this data in the discussion section.

Table 4.1.4.1 shows the distribution of ratings for each of the removed formats and the ranking of the formats by mean level of endangerment. Following the table are excerpts of the text participants used to explain their ratings for these file formats.

Rank	Format	Already Inaccessible	1-5 years	6-10 years	11-20 years	20+ years	Not Familiar	Mean	sd
1	**.mdb** Microsoft Access Database	0	1	1	0	2	6	2.25	1.45
2	**.wk1** Lotus 1-2-3 Worksheet	0	2	0	0	3	5	2.20	1.60
2	**.sdw** StarOffice Writer Text Document	0	1	1	1	2	5	2.2	1.45
3	**.swf** Shockwave Flash Movie	0	1	0	1	3	5	1.8	1.29
3	**.wri** Microsoft Write	0	1	0	1	2	6	1.8	1.29
4	**.spx** Ogg Vorbis Speex File	0	0	1	1	3	5	1.16	1.03
5	**.sgi** Silicon Graphics Image File	0	0	0	1	3	6	1.25	0.71

Table 4.1.4.1. Round 1 file format rating distribution, mean scores, and standard deviation for removed file formats, ranked by mean value.

1. .sgi, Silicon Graphics Image File, Version 0.97

Of all the seven removed file formats, the Silicon Graphics Image file format received the lowest mean rating score at 1.25, with the next highest mean score being 1.60 for the Ogg Vorbis Speex File. Only four of the ten participants knew enough about the format to rate it. Three of these participants rated it at 20+years. Reasons they supplied were simplicity (1 participant), available specifications (2 participants), and availability of rendering software (2 participants). The participant who rated the format at 11-20 years indicated shrinking support as the rationale for this rating.

2. .mdb, Microsoft Access Database, Version 7.0 for Windows

While only four of the ten participants rated the Microsoft Access Database format, it received the highest mean rating (2.25) of the seven removed formats. If this format had received this mean rating in Round 2, it would have been the second highest mean rating of all the formats. Three of the six participants who indicated that they were not familiar enough with the format provided their opinions on the format's accessibility. Most of these indicated that they believed the format is at risk. For example, one wrote, "I don't know this format very well, but given its nature I'd expect this to be a bit of a nightmare. Closed standard heavily dependent on a single vendor, complex, etc."

The participant who rated the format at 1-5 years wrote,

> Although this version of the Access format has been reverse-engineered it has never been clearly and formally documented, and Access has been the most difficult of the MS office formats to do this with. The code is the only real documentation. The other applications which can deal with the format (openoffice, libre office et al) are themselves large and complex software suites with many dependencies on other software environments which are unlikely to survive. MS themselves are not motivated to provide backwards compatibility for more than one or two versions.

One of the two participants who rated it at 20+ years also indicated, "This content is at risk. I am not comfortable putting a time limit on it. I do not think it will ever be inaccessible, it will just be harder to get to."

3. .spx, Ogg Vorbis Speex File, Version 1.1.12

The Ogg Vorbis Speex File format had a mean value of 1.16. Three of the five participants who rated it, rated it at 20+ years. Their rationale included the availability of specifications (2 participants) and the availability of rendering software (3 participants). The participant who rated it at 11-20 years indicated low uptake and ambiguity of specification quality as the rationale for these ratings. The participant who rated it at 6-10 years cited low adoption as a rationale for this rating.

4. .wk1, Lotus 1-2-3 Worksheet, Version 2.0

This file format also received a high mean rating (2.20) with five of the ten participants rating it. Two of the five participants rated it at 1-5 years, citing decline in commercial support (1 participant) and compromised functionality in available rendering software (1 participant). Two of the three 20+ ratings were from the two participants who indicated in each of their ratings that they believe access to content stored in file formats can be maintained indefinitely. The third of these cited the availability of open source software that can be used to render these files.

5. .wri, Microsoft Write, Version 1.0

Two of the four participants who rated this format rated it at 20+ years. Both of these participants provided non-descriptive answers stating that they believed there would be continued access to contents encoded in this format. The participant who rated it at 11-20 years wrote, “I couldn't track down either a specification or a modern open-source renderer, but the ubiquitous WordPad is apparently capable of opening this format and there have been some efforts to facilitate conversion." The participant who rated it at 1-5 years wrote, "Sounds early, sounds proprietary, doesn't look good."

6. .sdw, StarOfficeWriter Text Document, Version 5.0

There was a broad variation of answers among the five participants who rated this format. One of the two participants who rated the format at 20+ years indicated that there is open source software available to render files in this format. The three participants who rated the format at higher endangerment levels cited the decline in support from both commercial and open source communities.

7. .swf, Shockwave Flash Movie, Version 5

The Shockwave Flash file format had a mean rating score of 1.8, which is relatively high compared to the other file formats' mean scores. The one participant who rated it at 1-5 years cited incomplete specifications as the rationale for this rating. The participant who rated it at 11-20 cited its dependency on Adobe corporate support. One participant cited the availability of open source support as rationale for a 20+ years rating.

4.1.6. Factors Discussed in Comments

During the process of rating the file formats participants wrote short narratives to justify each of their ratings. Using the list of factors compiled for this study as a "start list," as recommended by Miles and Huberman (1994, p. 10), I coded the text of the justification text provided by participants. Almost all of the justification text conformed to this list of factors as participants naturally justified their ratings by discussing the factors that informed their decisions. I had also added one factor (value) to the list of factors the second group of Delphi experts rated in their second round of the study.

I had a second reviewer perform the coding task for ten file formats and calculated the inter-rater reliability coefficient to be 0.63. I met with the coder and discussed differences between our respective counts of each factor. We discussed each coding for each format and our respective rationale for our coding to identify divergences in our methodology. We identified several discrepancies that we would each address in a second round of coding:

- Discussions of XML schema documentation should not be coded as *specifications available*, but rather *technical dependencies*.
- Mentions of open source software should be coded as *legal restrictions*.
- Mentions of the format being readable in plain text should not automatically be coded as simplicity.
- The word standard should be evaluated more closely for whether it refers to *specifications available* or *standardization*.

After reviewing these discrepancies we each re-coded the first ten file formats and I recalculated the inter-rater reliability coefficient at 0.94. Keeping in mind the discrepancies identified through the coding process, I coded the text for the entire set of responses from the Round 1 format-rating questionnaire. Of the twenty-eight factors in the start list, eighteen appeared in the justification text. See *Table 4.1.6.1* for a list of all of the factors and the count of their appearance in the justification text.

Factor	Count
Rendering Software Available	162
Ubiquity	130
Specifications Available	111
Legal Restrictions	97
Complexity	63
Community/3rd Party Support	51
Specification Quality	46
Developer/Corporate Support	44
Standardization	42
Technical Dependencies	42
Rendering Software Feature/Functionality/Behavior Support	18
Backward/Forward Compatibility	12
Value	11
Compression	10
Lifetime	8
Ease of Identification	3
Technical Protection Mechanism	1
Domain Specificity	1
Expertise Available	0
Cost	0
Revision Rate	0
Geographic Spread	0

Factor	Count
Ease of Validation	0
Institutional Policies	0
Error-tolerance	0
Metadata Support	0
Storage Space	0
Viruses	0
Availability Online	0

Table 4.1.6.1. Appearance count of file format endangerment factors in justification text.

I marked text that mentioned the availability of software, applications, implementations, tools, viewers, web browsers, compilers, readers, and specific names of software applications that can render the formats rated as, *Rendering Software Available.* There were 162 instances of text that fell under this code, making it the most common factor mentioned.

The second most common factor was *Ubiquity*, with 130 instances in the text. I marked text that contained terminology like, “very widely used,” “ubiquitous,” “popular,” “common,” “widely adopted,” “heavy use,” “widespread,” “small” or “large market share,” “used extensively,” and citations of numbers of users as a reference to the presence or lack of ubiquity as a reported factor in participant ratings.

The factor, *Specifications Available* appeared 111 times in the format rating justification text. I counted terms and phrases like “open standard,” “documentation,” “schema,” “spec/specifications available, “open specifications,” “well documented,” "transparency of representation," and "lack of documentation" as indicators for this factor. It is important to note that there was some ambiguity around the term, "standard" in that in some cases participants used it to refer to the availability of specifications for a format, and in other cases they used it to indicate that it was adopted as a standard by standards issuing bodies such as the International Organization for Standardization (ISO) and the American National Standards Institute (ANSI).

The factor *Legal Restrictions* appeared 65 times in the file format rating justification text. I coded the text as *Legal Restrictions* whenever the word "open" was used in reference to software, specifications, and standards. I also coded text as such when participants used the phrases "freely available," "proprietary," "patent," "closed," "publicly available," "published/unpublished," and various forms of the word "license."

The *Complexity* factor appeared 63 times in the justification text. I coded the following phrases under this factor: "simple," "complicated," "complex," "basic," "straightforward," and some discussions of how the format is simple but may contain or link to more complex formats. There is some overlap between this last aspect of *Complexity* and the factor, *Technical Dependencies*, discussed below.

The factor *Community/3rd Party Support* appeared 51 times in the justification text. I coded certain references to "open source" software as *Community/3rd Party Support* when the file format being rated was not itself the open source software referenced by the participant. In addition to general references of "wide support," I also coded direct references to communities that support or have ceased support of a particular format under this factor. Lastly, any discussions of reverse-engineering a file format indicate a presence of a 3rd party developer and so I coded these instances under this factor.

The factor, *Specification Quality* is a sub-factor of *Specifications Available* that was recommended by a participant in the first round of the factor rating questionnaire. Instances of this factor co-occurred with the *Specifications Available* factor. I noted these instances only when specifications were discussed and qualified using phrases such as “well-documented,” “thoroughly described,” “good documentation,” “clear documentation,” “well understood,” “completeness is unclear,” “described in detail,” “complex specifications,” “it can be built from scratch using the documentation alone,” and “specifications… flawed,” The *Specification Quality* factor appeared in the text 46 times.

The factor, *Developer/Corporate Support* appeared in the justification text 44 times. I coded text under this factor when the support of the originating corporate developer was discussed in context of their continued support of the file format. Corporate developers that participants commonly mentioned in this regard were Microsoft, Google, Adobe, and Apple. I coded 42 instances of the factor, *Technical Dependencies* in the justification text. I coded under this factor text that indicated some kind of dependency on external resources such as other file formats, computing platforms, and software environments.

As mentioned previously, the term “standard,” was used in the justification text to refer both to specifications and the acceptance as a standard by a standards issuing body. I was careful not to code text as *Standardization* unless it was clear that it was not referring to specifications. I only coded it as such if a particular standards body was mentioned (W3C, ISO, ANSI). I coded a total of 42 instances of the *Standardization* factor.

There were 18 occurrences of *Rendering Software Feature/Functionality/Behavior Support*. Like *Specification Quality*, this is a sub-factor of a primary factor in the list, *Rendering Software Available*. This factor was not in the original list of coding factors; it surfaced as I re-considered some of the participant comments on the rendering software they discussed. I coded text under this new factor that discussed feature, functionality, and behavior support of the available rendering software for a particular file format. In most cases, participants discussed lack of feature/functionality, and behavior support as an influence in their format rating.

The factor, *Backward/Forward Compatibility* appeared in the text 12 times. I coded this factor when either backward or forward compatibility were specifically mentioned. Another factor that emerged from the format rating justification text was, *Value,* which appeared in the text 11 times. I coded the text as *Value* when participants wrote that they believed a great deal of valuable content was stored broadly in the format they were rating. Only one participant mentioned this factor at 11 different instances.

The *Compression* factor was mentioned 10 times in the text. I coded text under the *Compression* factor when participants specifically mentioned some kind of compression in their justification text. The factor, *Lifetime* was mentioned 8 times in the justification text. I coded text under the *Lifetime* factor when participants mentioned the lifespan of the file format, either generally or in specific years.

The last three factors that were represented in the justification text had relatively few appearances. *Ease of Identification* appeared three times, and *Technical Protection Mechanism* and *Domain Specificity* appeared only once. I coded text under *Ease of Identification* when a participant specifically wrote about the difficulty of identifying a format. The *Technical Protection Mechanism* factor appeared only once when a participant mentioned Digital Rights Management (DRM). I coded the one time a participant mentioned the specific domain with which a format is associated under *Domain Specificity*.

4.2. Questionnaire 3. Factor Rating.

In this section, I present the quantitative and select qualitative data collected through the factor rating questionnaire Delphi process. I present Spearman's rank correlation coefficient values for each factor, together with Round 1 and 2 mean scores and their calculated difference. I present mean scores for each factor rated in each of the Delphi rounds, and present Round 2 ratings and mean scores for each factor. I present a list of factors, ranked in order of the highest mean scores (most relevant) to the lowest mean scores (least relevant). Additionally, I present description and excerpts of participant justification text for each factor.

4.2.1. Delphi Termination Using Spearman's Rank Correlation

After the experts completed the second round of rating factors, I checked for answer stability for each question using Spearman's Rank Correlation (r_s). The calculated correlation coefficient values demonstrated that answer stability was reached for all file format ratings after the second round each item was rated. These calculations are shown in *Table 4.2.1.1.*

Original Factors	R1 Mean	R1 sd	R2 Mean	R2 sd	Mean Diff.	r_s
Backward/Forward Compatibility	1.14	0.50	0.59	0.54	-0.55	0.93
Community/3rd Party Support	1.14	0.67	1.05	0.69	-0.09	0.98
Complexity	0.96	0.52	0.68	0.60	-0.28	0.98
Compression	0.05	0.69	-0.05	0.52	-0.10	0.98
Cost	0.77	0.65	0.77	0.47	0.00	0.97
Developer/Corporate Support	0.32	0.75	0.50	0.77	0.18	0.97
Ease of Identification	0.68	0.75	0.59	0.83	-0.09	0.93

Ease of Validation	0.68	0.75	0.32	0.87	-0.36	0.88
Error-tolerance	0.50	0.89	-0.05	0.69	-0.55	0.93
Expertise Available	1.23	0.47	1.05	0.52	-0.18	0.95
Legal Restrictions	0.96	0.69	0.96	0.52	0.00	0.95
Lifetime	0.32	0.75	0.41	0.30	0.09	0.96
Metadata Support	0.14	0.67	-0.14	0.50	-0.28	0.98
Rendering Software Available	1.41	0.30	1.14	0.67	-0.27	0.96
Revision Rate	0.41	0.54	0.23	0.47	-0.18	0.98
Specifications Available	1.50	0.00	1.41	0.30	-0.09	0.99
Standardization	0.77	0.65	0.59	0.30	-0.18	0.97
Storage Space	-0.14	0.50	-0.23	0.47	-0.09	0.98
Technical Dependencies	0.86	0.67	1.05	0.52	0.19	0.95
Technical Protection Mechanism	0.32	0.75	0.32	0.75	0.00	0.97
Ubiquity	1.14	0.50	0.86	0.67	-0.28	0.96
New Factors	**R2 Mean**	**R2 sd**	**R3 Mean**	**R3 sd**	**Diff.**	r_s
Value	0.32	0.87	0.30	0.79	-0.02	0.89
Geographic Spread	-0.05	0.69	0.20	0.67	0.25	0.94
Domain Specificity	0.50	0.63	0.30	0.63	-0.20	0.97
Viruses	-0.23	0.47	-0.40	0.32	-0.17	0.96
Availability Online	0.46	0.69	0.40	0.57	-0.06	0.93
Institutional Policies	0.14	0.67	0.20	0.82	0.06	0.96
Specification Quality	0.86	0.67	1.00	0.53	0.14	0.98
Overall Mean Difference					-0.13	

Table 4.2.1.2. Factor means (where -0.50 is 'Not Relevant', 0.50 is 'Somewhat Relevant' and 1.50 is 'Very Relevant'), standard devaiations, means differences, and Spearman's rank correlation coefficient values.

4.2.2. Factor Rating Results

In this section, I present a table of data collected during the factor questionnaire Delphi process. This table, *Table 4.2.2.1*, shows the Round 2 distribution of ratings, means, and standard deviations for the 28 factors rated in this study, ranked in order of the highest mean scores (most relevant) to the lowest mean scores (least relevant). The highest rating, 1.41 (*Specifications Available*) indicates a relevancy rating between *somewhat relevant* and *very relevant*, but much closer to *very relevant*. The factors rated below 0.50 were removed from the list of factors presented to the special rater in Questionnaire 4 because anything below this rating level is less than *somewhat relevant*.

Rank	Factor	Not Relevant (-0.50)	Somewhat relevant (0.50)	Very Relevant (1.50)	Mean	sd
1	Specifications Available	0	1	10	1.41	0.30
2	Rendering Software Available	1	2	8	1.14	0.67
3	Technical Dependencies	0	5	6	1.05	0.52
3	Community/3rd Party Support	1	3	7	1.05	0.69
3	Expertise Available	0	5	6	1.05	0.52
4	Specification Quality	0	5	5	1.00	0.53
5	Legal Restrictions	0	6	5	0.96	0.52
6	Ubiquity	1	5	5	0.86	0.67
7	Cost	0	8	3	0.77	0.47
8	Complexity	1	7	3	0.68	0.60

Rank	Factor	Not Relevant (-0.50)	Somewhat relevant (0.50)	Very Relevant (1.50)	Mean	sd
9	Standardization	0	10	1	0.59	0.30
9	Backward/ Forward Compatibility	1	8	2	0.59	0.54
9	Ease of Identification	3	4	4	0.59	0.83
10	Developer/Corporate Support	3	5	3	0.50	0.77
11	Lifetime	1	10	0	0.41	0.30
12	Availability Online	2	7	1	-0.06	0.57
13	Technical Protection Mechanism	4	5	2	0.32	0.75
13	Ease of Validation	5	3	3	0.32	0.87
14	Value	4	4	2	0.30	0.79
14	Domain Specificity	3	6	1	0.30	0.63
15	Revision Rate	3	8	0	0.23	0.47
16	Geographic Spread	4	5	1	0.20	0.67
16	Institutional Policies	5	3	2	0.20	0.82
17	Compression	6	5	0	-0.05	0.52
17	Error-tolerance	7	3	1	-0.05	0.69
18	Metadata Support	7	4	0	-0.14	0.50
19	Storage Space	8	3	0	-0.23	0.47
20	Viruses	9	1	9	-0.40	0.32

Table 4.2.2.1. Round 2 factor rating distribution and mean scores.

4.2.3. Justification Text

This section contains selections from the justification text provided by participants during the Delphi rating process. Most of the text represented here is from Round 2 of the Delphi, and was selected from Round 1 answers when participants referenced their Round 1 justifications in their Round 2 explanation. This was particularly common when participants indicated that they had not changed their ratings in the second round. Not all participants consistently provided in-depth or very descriptive justification text. The experts I provide are the best representations of particular themes that emerged in the text.

1. Backward/Forward Compatibility - *whether or not newer versions of the rendering software can render files from older versions, or whether or not older versions of rendering software can render files from newer versions.*

The *backward/forward compatibility* factor shared the rank of 9th with *standardization* and *ease of identification* with a mean rating of 0.59. This rating indicates that, on average, it was considered to be "Somewhat relevant." Eight of the eleven respondents indicated it was somewhat relevant, where two rated it at very relevant and one participant indicated that it was not relevant at all.

Several participants noted that they felt *backward compatibility* and *forward compatibility* should be separate factors. These participants indicated differing opinions as to which of the two was more relevant, though most participants agree that *forward compatibility* is less relevant. As one participant wrote, "Backward compatibility is very relevant (as an indication for technology watches that the software support is declining, as a warning that older versions of software will need to be maintained to access the content), while forward compatibility is not relevant (do we really expect older versions of software to continue to be updated with support for newer formats - seems unrealistic).” Another wrote, “Forward compatibility is more important because it allows old files to be accessible with the current standard systems. Backward compatibility is somehow relevant, in the sense that it takes some time until the great majority of the systems are brought to the newer standard.”

Measurement: Four participants provided recommendations on how to collect data for this factor. One recommended reviewing case histories for the selected format. This participant wrote the same or a similar suggestion for a majority of the remaining factors. The other three indicated that the factor could be measured by noting the number of versions of a particular file format and the number of previous versions an application supports. One participant wrote, “To measure, groups of compatible software versions need to be defined. See word 97-2003 for example. In software development there is a naming convention for numbering the versions: see major & minor version numbers: http://en.wikipedia.org/wiki/Software_versioning”

2. Community/3rd Party Support - *the degree to which communities and/or parties beyond the original software producers support the file format.*

Seven participants rated the *community/3rd party support* as being very relevant as a cause of file format endangerment. This factor shares the ranking of third highest relevancy rating with a mean score of 1.05, with *technical dependencies* and *expertise available*. One participant who rated this factor as *very relevant* wrote, “When there is lot of third party/community support, we are less vulnerable to the risk of a single point of failure (the original maintainer ceasing to support it).”

Measurement: Five participants shared their thoughts on how data could be collected for this factor. One participant recommended measuring message board engagement and market share of 3rd party support. Similarly, another participant recommended measuring the number of 3rd party applications. Another participant wrote of the difficulty of measuring this factor: “Like the one above, this is hard to measure confidently for the same reasons - (1) how do we define support? (2) what can we rely on without objective testing? There are many different ways a format could be 'supported' and some are more valuable than others for ensuring long-term preservation.”

3. Complexity - *relates to how much effort has to be put into rendering and understanding the contents of a particular file format.*

The *complexity* factor received a mean rating score of 0.68. This places it just above the *somewhat relevant* level. The one participant who rated it as *not relevant at all* wrote, “Complexity has to be addressed ’today‘ in order to create and use certain formats in the first place. In a sense, preservation activities inherit the effort to render-for-access-and-use-today.” The other participants discussed how higher levels of complexity in a file format increases the difficulty with which it can be validated, migrated, and with which rendering software can be developed to access it.

Measurement: Four participants provided input on how to measure this factor. One participant wrote, “# of pages in specification, # of formulas in specification?, number of features supported by the format, depending on the content type -- # of supported color spaces, etc.” Another suggested using function point analysis: “As measures, one could use software metrics if appropriate or derive models similar from function point analysis: http://en.wikipedia.org/wiki/Function_point.”

4. Compression - *whether or not, and the degree to which a file format supports compression.*

The factor *compression* received one of the lowest relevancy ratings at -0.05. Six of the eleven respondents rated it at *not relevant at all*, where the remaining five rated it at *somewhat relevant*. Three participants who rated it as *not relevant at all* indicated generally that compression did not act as a cause of file format endangerment, one participant wrote more specifically that compression affects access to digital contents on a file level, but not necessarily the endangerment of a specific file format. One participant wrote, “Compression has an impact on the files, and could make them unrenderable. However, is not necessarily a cause to file format endangerment.” One participant countered, “I can understand the opinions of those who say that compression itself isn't particularly relevant, but given that there are proprietary compression algorithms loose in the world, I'm not willing to say it's never relevant. Having data in a proprietary, compressed format and then watching the vendor who created the format go under leaves you with a serious problem. So, compression may only be relevant in some situations, but definitely can be relevant.”

Measurement: Only one participant provided information on how to measure this factor. This person suggested counting “the number of supported compression algorithms. The number of alternative compression schemas (configurations).”

5. Cost - *The cost to maintain access to information encoded in a particular file format, e.g. to migrate files, to maintain the rendering software, or to run an emulation environment*

The *cost* factor had a mean rating that was slightly more than *somewhat relevant* at 0.77. One participant pointed out in her comments that they believed other participants misinterpreted the question in Round 1: "Again the description (cost to migrate, run an emulation env., etc.) and the comments show that people are answering the wrong question (To what degree does the format's cost affect the ability to preserve) instead of the question I think we're supposed to be answering (To what degree does the format's cost lead to the format becoming endangered). I don't think the user take-up and software production is concerned at all with the cost to migrate, emulate, etc. Unfortunately preservation difficulties aren't much of a factor in user and tool adoption!"

Two of the three participants who rated it as *very relevant* indicated that higher costs to maintain access to contents of a file in a given format would increase endangerment, but did not elaborate further. One of these three wrote, "high costs are in and of themselves a direct push towards format obsolescence, because they will push people to find something else to use." Two of the eight participants who rated it as *somewhat relevant* indicated that the question of assessing cost is very complicated. One wrote, "The costs are very important in general, but this still needs to be considered in conjunction with the complexity, financing and revenue. The measures including the context factors are hard to be computed, a much simpler indirect measure would be popularity."

Measurement: Other than the one participant's repeated suggestion of analyzing case histories, two other participants provided suggestions on how to collect data for this factor.

One suggested: “Some KO [Knowledge Organization] criteria to identify imminent risks could be defined: e.g. high cost for current software/file format while free, stable and popular open source solutions exist.” The other participant suggested something similar: “You could measure this with a TCO [Total Cost of Ownership] model and run the costs through the business processes needed to maintain and provide access to the files.”

6. Developer/Corporate Support - *whether or not the entity that created the original software that produces output in the file format continues to support it.*

The *developer/corporate support* factor received a mean rating of 0.50, and though five participants rated it as *somewhat relevant*, half of the remaining six participants rated it as *very relevant*, and half as *not relevant at all*. Those who rated it as not relevant at all indicated that whether or not the original developer/corporation continued to support the format had no bearing on whether or not the file format would have the support it needs to remain accessible. As one participant stated, “The fact that the company which originally produced software that produced output in a format no longer supports it just isn't that big a deal assuming that other software developers have picked up the format and *are* supporting it.”

Two of the five participants who rated it as *somewhat relevant* indicated that the effect of diminished developer/corporate support on the endangerment level of a file format was contingent on whether or not the format was proprietary and/or had support from 3rd party developers. Two of the three participants who rated it as *very relevant* indicated that especially for proprietary formats, continuing support from the original developer is very important for maintaining access to the contents stored within them.

Measurement: One participant suggested that information could be collected about the original supplier about whether or not they continue to support the format.

7. Ease of Identification - *the ease with which the file format can be identified.*

The factor, *ease of identification*, had a mean value of 0.59, and shared this rating with *standardization* and *backward/forward compatibility*. While it was rated solidly at the *somewhat relevant* level, the ratings were spread fairly evenly among the choices. There were three participants who rated it as *not relevant at all*. One of these participants clearly stated that whether or not a file format can be identified is not a cause of file format endangerment. Another similarly wrote, “There are a number of file formats that simply don't lend themselves easily to signature-based identification methods, and that does not stop people from using them and I don't see it as contributing directly to their obsolescence.”

One participant agreed with this rationale, but rated this factor as *somewhat relevant* and wrote, “presumably well understood and popular formats are supported by identification tools so it can be a measure of how well its supported.” Another participant justified this rating by writing, “Identification is very important in order to find the right tool for rendering a format.” Three of the four participants, who rated it as *very relevant*, noted similar reasons as those who rated it as s*omewhat relevant*. One however, while rating it as *very relevant* for preserving a file in a given format, expressed uncertainty over whether it was actually a cause of file format endangerment. In summary, participants believe that file format identification is very important for preservation activities, but it does not have a direct effect on file format endangerment.

Measurement: One participant recommended measuring the number of tools available to identify the particular file format. Two other participants suggested collecting this kind of data in a registry.

8. Ease of Validation - *the ease with which the file format can be validated, where validation is the process by which a file is checked for the degree to which it conforms to the format's specifications.*

While the ratings of the factor *ease of validation* were spread out among the choices, five of the ratings fell under *not relevant at all*, resulting in a final mean value of 0.32. Similar to the factor *ease of identification*, three of the five participants who rated this factor as *not relevant at all* stated that it was useful for preservation purposes, but it was not an actual cause of file format endangerment.

Two participants who rated it as *somewhat relevant* and *very relevant* largely justified their ratings in terms of preservation actions. For example, one of these participants wrote, "It is important to understand if a file is a valid instance of what it purports to be. If the file is not valid then a future migration strategy / action may not succeed. " Five of the six participants who rated it as either somewhat relevant or very relevant discussed the fact that validation is difficult for many file formats.

Measurement: One participant suggested referencing a registry, if one existed, that contained information on validation methods as a source of data to measure.

9. Error-tolerance - *the degree to which this format is able to sustain bit corruption before it becomes unrenderable.*

With a mean value of -0.05, the *error-tolerance* factor shares the fourth lowest rating rank with *compression*. Two participants who rated it as *not relevant at all* stated that the factor affects access on the file level, but not the file format in general. The five other participants who rated it as *not relevant at all* indicted that while low-error tolerance is good for accessing digital content, it has little effect on file format endangerment.

The one participant who rated it as very relevant acknowledged this as well: “This is very relevant, but not so much from format endangerment, but instability issue. Note that this is more about the integrity of the infrastructure of a preservation programme than the format itself. Bit level preservation doesn't really care about file format.” Those who rated it as somewhat relevant acknowledged the existence of file formats that are susceptible to errors and the fact that this may have a secondary impact on a file format’s endangerment level. One participant wrote, “This feature may encourage the use of a specific format and therefore keep it 'alive'. It does not have a major impact on the format endangerment.”

Measurement: There were no suggestions for measuring this factor.

10. Expertise Available - *the degree to which technological expertise is available to maintain the existence of software that can render files saved in this format.*

The factor, *expertise available*, shares the third highest mean value ranking with *community/3rd party support* and *technical dependencies*. No participants rated it as *not relevant at all*, five rated it as *somewhat relevant*, and six rated it as *very relevant*. Those who rated it as very relevant indicated that it is very important to have people available with the technological knowledge to create and maintain the tools necessary to maintain access to file format content. In particular, one participant wrote, “We have to have people who understand the format to be able to write and maintain associated software.” Similarly, another wrote, “Technical experts are needed to understand and render files, write software and maintain systems. You need good technical people not only for software but also to do things like reverse engineering and, even, to read and analyze specification documents.”

Participant justifications varied for those who rated this factor as somewhat relevant. Two participants felt that this factor could be rolled into the *community/3rd party support* factor. One participant felt that while it is a relevant factor, it is only relevant in the limited instances when a file format is not ubiquitous or well known. Another participant expressed questions about the meaning of “expertise available,” writing,

> Still think this is a highly contextual issue that's difficult to make blanket statements on. But another issue here is what do we mean by ’expertise available.’ The OAIS notion of representation information is at least in part a way of planning for the day when *experts* aren't available, and someone needs to re-educate themselves on how a format stored data in order to try to decipher a file. So, if we expand "expertise" to include "recorded expertise," I'd probably say 'very relevant' but if it's "there's a human who knows this stuff already available" then I think it's a bit less relevant.

Measurement: One participant wrote, “Measurement: the expertise/knowledge of the IT-people Collection of data: tests, hackathons etc.” Another suggested, “This could be measured by how many IT or other related areas service this (e.g. COBOL programmers employed if COBOL was a format).”

11. Legal Restrictions - *the degree to which this file format is or can be restricted by legal strictures such as licensing, copy and intellectual property rights.*

The factor *legal restrictions* had a mean rating value of 0.96. Like *expertise available*, this factor received no *not relevant at all* ratings. Six participants rated it as *somewhat relevant,* and five rated it as v*ery relevant*. Both participants who rated this factor as *very relevant* and *somewhat relevant* indicated that legal restrictions could impede the ability to develop tools to maintain access to the format. One participant who rated it as somewhat relevant noted, "For applications where a repository wants to employ emulation as a way to provide access to objects over time, this is an issue."

Two participants who rated this factor as *somewhat relevant* noted its connection to the availability of specifications. One participant wrote,

> Here is an element that, if not a cause, has endangerment as a consequence. I'd like to see a highlighting or emphasis on the availability of documentation or specifications. This does not mean free: we can probably all afford the cost (medium high) for ISO, ITU, NISO, and SMPTE standards documents, and they do provide needed information. In contrast, some industry specifications (e.g., the Red Book that governs Compact Disc Digital Audio, at one point listed at $5,000) are very pricey. Worst is the status for some proprietary formats, where the specifications seem not to be for sale.

Measurement: One participant recommended, "It can be measured by # of patents, existence of different legal restrictions, the license." Another similarly wrote, "You could evaluate this based on the terms surrounding the file format." Another noted, "The PREMIS metadata standard has semantic units for capturing this, that might need to be extended. The EU project 'KEEP' has a lot of case studies on this topic."

12. Lifetime - *the length of time the file format has existed.*

The *lifetime* factor received the relatively low mean rating value of 0.41. This places it just below the *somewhat relevant* rating. While most participants indicated that it was somewhat relevant as a cause of file format endangerment, one participant rated it as not relevant at all. In the first round, one participant rated it as *very relevant*, and four rated it as *not relevant at all*.

Four participants stated that the *lifetime* factor was not a strong indicator of file format endangerment. One participant wrote, “Active lifetime of a format can be related to uptake - the longer lived it is and is still being actively used then it is likely that there will be community support and other tools. So it is an indicator of use rather than a direct cause.” The participant who rated it as not relevant at all wrote, “How long something has been around isn't an indicator as to whether it will still be there next week or not.”

Measurement: There were three suggestions to reference the creation dates and how long it has been supported. One participant suggested that this data can often be found in existing file format registries.

13. Metadata Support - *whether or not the file format allows for the inclusion of metadata.*

The *metadata support* factor has the second lowest mean value of -0.14. Seven of the eleven participants rated it as *not relevant at all,* four rated it as *somewhat relevant* and zero participants rated it as *very relevant*. Four participants who rated this factor as *not relevant at all* indicated that they did so because, although metadata is useful in understanding more about a particular file, it is not a cause of file format endangerment. One participant wrote, “I don't think this is a cause of endangerment - in the long run it might be a reason why certain file formats become more used, but I don't think that this is a reason why someone might choose one file format over another."

All four participants who rated it as *somewhat relevant* indicated that metadata is useful to render files in the long-term, but its absence does not necessarily cause the file to be endangered. One participant wrote, "The metadata may carry important preservation information like provenance, identification, fixity. The technical and editorial metadata is very important for preserving the content, but it is not mandatory to be embedded in the content file. Still embedding critical metadata is the more robust solution on ensuring the synchronization with the content." Similarly, another wrote, "Hard to see this as a cause for endangerment. Some metadata (of a technical sort) is need[ed] to play a file, today and tomorrow. Other (descriptive and/or administrative) metadata is a desirable element, with its own long-term value, but its absence does not endanger the file."

Measurement: One participant suggested, "Measurement: characteristics of the file format Data collection: documentation of the format," and another, "Data collection: documentation of the format."

14. Rendering Software Available - *whether or not any type of software is available that can render the information stored in this file format.*

The *rendering software available* factor had the second highest mean value ranking at 1.14. One important change to note is that in the second round of rating, one participant created a causal model with the listed factors, wherein the participant defined file format endangerment as the lack of rendering software, in lieu of the definition I provided. Based on this model, this participant changed his response from *very relevant* to *not relevant at all*.

Seven of the eight participants who rated this factor as *very relevant* indicated that the availability of rendering software was necessary for accessing content stored in particular file formats. This is noted in the comment, “If you can't see or use the content, then the file is not usable and is therefore the file format is past endangered into obsolete!” Another participant commented, “Certainly in the case of proprietary formats rendering software is essential. In case of open formats, e.g. pdf, less.” Of the two participants who rated the factor as *somewhat relevant*, one commented, “The more the better…useful as a contribution to a larger judgment of sustainability.”

Measurement: One participant noted, “A well-developed format registry could list the applications that will render a format.” Another wrote, “Easily measured by trying to render it.”

15. Revision Rate - *the rate at which new versions of this file format’s originating software are released*.

The revision rate factor received a low mean rating score of 0.23. This places it between *somewhat relevant* and *not relevant at all*, but closer to *somewhat relevant*. Eight participants rated it as *somewhat relevant* and three as *not relevant at all*. Two of the three participants who rated it as *not relevant* indicated that they did not believe it had an effect on file format endangerment. Specifically, one participant wrote, “Revision rates *in and of themselves* simply don't impact a file format's viability and longevity. If there's a lot of *incompatible* revisions, then sure, that's a problem, but that's a different issue." Another explained, "A vendor's decision on release cycle's doesn't tell you anything about the potential longevity of the vendor. And what preservationist has to care about is whether the organization producing the software is going to be around and still supporting the software."

Five of the eight participants who rated the factor as *somewhat relevant* noted that this factor was an indicator of the stability or volatility of the format that three of whom indicated was an indirect cause of file format endangerment, contingent on backward/forward compatibility. One participant noted, "It is an indicator of format stability. The software and the processes need to keep the pace with the revision specific changes. High revision rate might not be a problem if many versions are backward and forward compatible." And another similarly stated, "Could make it more endangered if it indicates file format volatility. It could also be more endangered, based on renderability (so if there was backward compatibility, then it could harder to render different versions of the same file format)."

Measurement: One participant suggested, "It can be measured by # of years between revisions, average length of time for each revision." Another said, "Revision rate can be computed per year probably by computing the overall value, but also by grouping on backward/forward compatibility"

16. Specifications Available - *whether or not documentation is freely available that can be used to create or adapt software that can render information stored in this file format.*

The *specifications available* factor received the highest mean ratings value of all the factors, at 1.41. There were ten ratings of v*ery relevant* and one rating as *somewhat relevant*.

Four participants cited the importance of having specifications available for the continued development of rendering software and supporting tools. One participant wrote, “Can help development of a renderer if one does not exist.”

Another participant explained, “Being able to know what the format should look like and how it should interact with rendering software enables someone else to recode to enable the format to survive.” This comment is a clear statement of the importance of having specifications to aid understanding of how content stored in a particular file format should appear when rendered properly. The second participant who noted this importance wrote, “The value of specifications lies less in being able to create new or adapt old software in order to render a format, but in the information they hold which tells us what a format should look like.” The one participant who rated this as somewhat relevant did not provide an explanation.

Measurement: The two participants who made recommendations for measurement suggested conducting general searches online and engagement with communities for available specifications.

17. Standardization - *whether or not this file format is recognized as a standard for use and/or preservation by a reputable standards body.*

With a mean rating value of 0.59, the *standardization* factor was rated slightly higher than *somewhat relevant*. Ten participants rated it as *somewhat relevant* and one rated it as *very relevant*. Seven of the nine participants who rated it as *somewhat relevant* acknowledge the impact that standardization has on the popularity and support of a particular file format, but that lack of standardization does not impact a file format's level of endangerment. On this topic one participant wrote,

> Standardisation shows that, at some point in time, there was community support and some specifications. It depends on the age and level of take-up as to whether it is a cause of endangerment. However I don't think file formats which are not standards are more likely to become endangered. It is about whether they are fit for purpose and preservable as to whether a format survives (or able to be read and too expensive to move).

The one participant who rated this factor as *very relevant* did not provide rationale for this rating.

Measurement: One participant noted, "Whether a standards body has endorsed a specification should be easy enough to measure (yes/no)." Another recommended that information on standardization could be found by searching for publications of standards by reputable standards bodies.

18. Storage Space - *the average amount storage space a file saved in this format requires when saved.*

The *storage space* factor shared the second lowest mean score rank with the *error-tolerance* factor and the *metadata support* factor, with a mean rating value of -0.23. Eight of the eleven participants rated it as *not relevant at all* and three rated it as *somewhat relevant*. Four of the eight participants who rated it as *not relevant at all* stated that they did not believe it was a cause of file format endangerment. For example, one wrote simply, "Hard to see this as a *cause* of endangerment," and another wrote only, "the amount of storage space does not seem relevant at all."

One of the three participants who rated this factor as *somewhat relevant* wrote, "If all things are equal and a particular format takes more space (thus is more expensive to store) than theoretically it could lead to a decline in usage and support." Another indicated that this would be relevant if "cost was a driver."

Measurement: Participants did not provide measurement suggestions for this factor.

19. Technical Dependencies - *the degree to which this file format depends on specific software, operating systems, and hardware in order for its contents to be successfully accessed or rendered.*

The *technical dependencies* factor received the mean rating value of 1.05, as did, *community/3rd party* support and *expertise available.* Two of the six participants who rated it as *very relevant* noted that the greater dependency a file format has on various technological environments, the more difficult it is to maintain access to the contents stored within it.

Participants who rated it as s*omewhat relevant* presented the same rationale as participants who rated it as *very relevant*. One participant wrote, "The more dependent a format is on a particular technological frame, the greater the necessity of maintaining some or all of that frame to insure continuing access. And that is the difference between maintaining a file, and having to maintain a whole machine." Two participants, one who rated it as *very relevant* and one who rated it as *somewhat relevant* indicated that the occurrence of technical dependencies is not very common. One of these participants wrote, "If the file format is tightly tied to a specific environment, or even a proprietary format for a technical instrument, then if the technical dependency becomes obsolete, then there needs to be more work put into keeping the format alive. However I'm not sure that this is all that frequent in real life."

Measurement: One participant suggested, "You could measure this by the amount of a technical stack is needed to render the file." Another wrote, "Measurement: an adequate description of those technical dependencies Data collection: appropriate metadata about the file format such as in a file format registry."

20. Technical Protection Mechanism - *whether or not this file format allows for or is encumbered by technical protection mechanisms such as Digital Restrictions Management (DRM).*

The *technical protection mechanism* factor received the low mean rating of 0.32, which it shares with *ease of validation*. This places it below the *somewhat relevant* rating. There was a wide spread of responses, with four participants rating it at *not relevant at all*, five at *somewhat relevant*, and two at *very relevant*.

Of the two participants who rated it as *very relevant*, only one of them provided justification for the rating. This person wrote, "TPMs [Technical Protection Mechanisms] are a forced, artificial technological dependency which has been specifically designed to be difficult to circumvent or provide technological substitutes to insure on-going renderability. TPMs automatically consign a file format to a short life span as far as I'm concerned."

Of the five participants who rated it as *somewhat relevant,* three indicated that the access restrictions created by technical protection mechanisms prohibit continued access to the contents encoded in the format. They did not explicitly state that the access restrictions were a cause of file format endangerment.

Three of the four participants who rated it as *not relevant at all* acknowledged that while technical protections mechanisms prohibit access to content on a file level, they are not necessarily associated with particular file formats, and therefore do not affect endangerment levels of a particular file format. One participant wrote, “The capability to protect is not a cause for endangerment. Protected files, however, are a problem.” Another participant explained, “Typically, the digital restrictions are not mandatory to be embedded within the file format. They are used in particular contexts, where there is a strong reason to apply them and they are implemented at application level. Digital rights are encoded in metadata out of the content file, in the most of the cases.”

Measurement: One participant wrote, “Measurement: metadata about the format Data collection: via file format registry.”

21. Ubiquity - *the degree to which use of this file format is widespread and in common use.*

The factor, *ubiquity*, received a mean rating value of 0.86; between *somewhat relevan*t and *very relevant*. Five participants rated it as *very relevant*, five as *somewhat relevant* and one as *not relevant at all*. Participants who rated it as *very relevant* cited several different reasons. One participant wrote, “This is the opposite of obsolete!” Another wrote, “Widespread and common use means a viable market for rendering software and longevity for the format.” Overall, those who rated it as *very relevant* indicated a connection between ubiquity and continued support for the format.

Those who rated it as *somewhat relevant* indicated the same or similar rationale as those who rated it as *very relevant*, but included additional qualifiers to their statements. For example, one participant wrote, “The assumption is that if widely used the format will be less likely subject to obsolescence. Depends however also on things like the viability of the supplier, whether it is proprietary or not and the emergence of new more interesting formats.” And another stated, “In a utopian world there would be a small number of ubiquitous formats. This is not the case now, and is unlikely to be the case in the future.”

The one participant who rated this factor as *not relevant* said that there is not a good way to measure it. This participant wrote, “In terms of a cause for endangerment, I would prefer to say that this factor sets a context for risk assessment. But I despair of a solid metric for expressing ubiquity.”

Measurement: One participant provided the following suggestions for measurement: “Measurement: the number of files, the use of software that produced this format Data collection: web search, analysis or count of files with that format.” Another suggested consulting PRONOM.

4.2.4. New Factors

This section presents mean ratings and justification text for the factors that were suggested by participants in Round 1 of the factor rating Delphi process, shown in *Table 4.2.4.1*. These factors were rated for the first time in Round 2 and a second time in Round 3 of the factor rating Delphi process. Note that only one of the factors, *specification quality*, received a mean rating value above 0.50. This indicates that specification quality was the only factor participants considered to be above the *somewhat relevant* rating.

Factor	Not Relevant (-0.50)	Somewhat relevant (0.50)	Very Relevant (1.50)	Mean
Value	4	4	2	0.30
Geographic Spread	4	5	1	0.20
Domain Specificity	3	6	1	0.30
Viruses	9	1	0	-0.40
Availability Online	2	7	1	0.40
Institutional Policies	5	3	2	0.20
Specification Quality	0	5	5	1.00

Table 4.2.4.1. Ratings of new factors.

1. Value - *the degree to which information encoded in this format is valued.*

The factor, *value*, received a mean rating value of 0.30, placing it below the *somewhat relevant* rating. The ratings were distributed widely across the three choices with two participants rating it as *very relevant*, four as *somewhat relevant*, and four as *not relevant at all*. The two participants who rated it as very relevant indicated that the value of the content stored in a file format had a direct effect on the format's level of endangerment. One of these participants wrote, "One of the fundamental reasons for preserving digital content is because it is of inherent value to an organization, individual or society at large. The custodian of the information is responsible for determining its value." The other participant who rated it as very relevant wrote, "If the information is not valuable the file format is more prone to endangerment."

Three of the four participants who rated this factor as *somewhat relevant* indicated that the value of the content stored in a particular file format indirectly influences whether or not there is human motivation to take actions to maintain access to the format. One participant wrote, “Value may indirectly & positively affect incentives to maintain or create rendering software for the format; the increased likelihood of rendering software is a decreased likelihood of extinction; hence decrease in endangerment.”

Those who rated it as *not relevant at all* noted that since there was no direct relationship between the value of the content and the file format’s endangerment level, it should not be considered a relevant factor. One participant noted, “The value of data qua data is separate and distinct from the format it's stored in, and does not affect a format's viability one way or another.” Another participant similarly wrote, “This is about the content not the file format as such. I don't see an immediate relationship.” Two of the four participants who rated this factor as *not relevant at all* did mention that value could motivate preservation action, though they did not consider it a relevant factor as a cause of file format endangerment.

Measurement: Participants did not make suggests on how to measure this factor.

2. Geographic Spread - *the way in which a file format is spread across the world; whether spread thinly across the globe or condensed heavily in a particular area.*

The *geographic spread* factor’s mean rating value was 0.20. Four of the ten participants rated it as *not relevant at all*, five at *somewhat relevant*, and one at *very relevant*. The one participant who rated this factor as *very relevant* wrote of this rating, “If a file format is spread thin or heavily condensed, other parties might not be interested in preserving the format. An example that I know of is a particular, only locally used, file format used by a partner organisation which was deemed ’too obscure’ by PRONOM and thus was not added to their file format registry.”

Seven participants who rated this factor as *somewhat relevant* and not relevant at all noted that geographic spread is a proxy for the factor, *ubiquity* or what two participants referred to as *adoption*. One participant who rated it as *somewhat relevant* suggested that it might be easier to measure than the more general ubiquity factor.

Measurement: Participants did not make suggestions on how to measure this factor.

3. Domain Specificity - *the degree to which the format is used only within specific domains.*

The *domain specificity* factor received the low mean rating value of 0.30. One participant rated it as *very relevant*, six as *somewhat relevant*, and three as *not relevant at all*. The participant who rated it as *very relevant* wrote, “If the format is used only within specific domains this might be an indication for endangerment.” Participants who rated it as *not relevant at all* indicated that other factors already discussed covered the aspects of file format endangerment that *domain specificity* addresses.

Those who rated this factor as somewhat relevant had many different rationales. One participant indicated that domain specificity could be detrimental to a file format’s long-term accessibility and one indicated that it could be beneficial. Two participants acknowledged both the positive and negative affects this factor could have on file format endangerment.

Measurement: Participants did not make suggests on how to measure this factor.

4. Viruses - *the degree to which the format is susceptible to containing or being damaged by viruses.*

With a mean rating score of -0.40, the *viruses* factor received the lowest mean rating score of all the factors. Nine of the ten participants rated it as *not relevant at all* and one participant rated it as *somewhat relevant*. Eight of the ten participants who rated this factor as *not relevant at all* indicated that their rating was based on the fact that viruses are not endemic of any particular file format and therefore could not be a relevant factor for indicating file format endangerment for individual file formats. One participant wrote,

> Everything's susceptible to being damaged, so, that's not an issue. And honestly, given the range of data-based exploits available today I don't think the possibility of containing a virus is really that much of a contributor to a format being endangered. Unless someone can persuade me that the digital preservation world is going to abandon self-extracting zip and 7z files and image formats shown to host exploits like TIFF and BMP, I think none of us really believe that virus vulnerability affects a format's longevity.

The one participant who rated this factor as somewhat relevant indicated uncertainty that viruses can affect particular formats over others. This participant wrote, “But . . . are there really differences? Do some format[s] have greater susceptibility? I don't know.”

Measurement: Participants did not make suggests on how to measure this factor.

5. Availability Online - *the degree to which files in this format are available on the Web.*

The factor, *availability online*, received the second highest mean rating value of all of the new factors, but at 0.40, participants overall did not consider it to be relevant as an indicator of file format endangerment. Seven participants rated it as *somewhat relevant*, two rated it as *not relevant at all*, and one participant rated it as *very relevant*. The one participant who rated it as very relevant wrote, “If the availability is wide, this lessens the endangerment.”

The seven participants who rated it as *somewhat relevant* provided various reasons. One participant wrote, “Availability online may indirectly & positively affect the diversity of users; the number of rendering clients; and the value; all of which may indirectly increase incentives to maintain or create rendering software for the format, or make such incentives less likely to be affected by changing economic conditions; the increased likelihood of rendering software is a decreased likelihood of extinction; hence decrease in endangerment.” Another wrote, “I think one of the other reviewers put it best when they said ‘Although the web is a ubiquitous mechanism for publishing and sharing content, it is certainly not the only environment for rendering and providing access to digital content.’ And digital preservationists take it as a given that they have to collect representation information, and getting it through the post is just as valid a means as downloading.”

Overall, participants appeared to be working with different interpretations of the factor. One participant noted, “It is clear from the answers to the last questionnaire that we all had different interpretations of what this means! If this means that information about the format is more easily available, then that reduces endangerment. I'm not sure that being able to render an object using a web browser helps.”

Measurement: Participants did not make suggestions on how to measure this factor.

6. Institutional Policies - *the degree to which a file format is affected by institutional polices, such as whether or not an institutional policy states that content encoded in this format will be collected and preserved.*

The factor, *institutional policies*, was rated very low with a mean rating value of 0.20. While two participants rated this factor as *very relevant*, three rated it as *somewhat relevant*, and five rated it as *not relevant at all*. The two participants who rated it as *very relevant* indicated that institutional policies play an important role in digital preservation. One participant wrote, “In a robust preservation strategy it is the specific institutional policies around the acceptance and long term management of file formats that should guide and identify any risks to file format endangerment. I see this as the key overarching factor, that should be defined and maintained on a regular basis” The other participant wrote, “If an institute makes an effort in preserving the format, the danger will be less.”

The three participants who rated this factor as *somewhat relevant* indicated that this could be an indirect factor that affects file format endangerment. One participant wrote, “Formalization in institutional policies may indirectly increase incentives to maintain or create rendering software for the format, or make such incentives less likely to be affected by changing economic conditions; the increased likelihood of rendering software is a decreased likelihood of extinction; hence decrease in endangerment.”

Those who rated this factor as *not relevant at all* presented various viewpoints on their ratings. Four of the five participants indicated that policies were a reaction to format risk and did not cause it. One participant who wrote, “If a format is widely used, it may influence the institutional policy, but I don't think it will work the other way around.” One participant did not believe that institutional policies could not have a strong effect on file format endangerment. This person wrote, “just not seeing how institutional policies influence a file format's viability. Unless it's at the level of 'thousands of institutions world-wide are forbidding the use of this format for preservation which is leading people to abandon it,' in which case, we're back to an adoption issue, not a policy issue.”

Measurement: Participants did not make suggestions on how to measure this factor.

7. Specification Quality - *(sub-factor of "Specifications Available") the understandability and usefulness of the format's available specifications in maintaining access to content encoded in that format.*

I presented this last factor, *specification quality*, to participants to rate as a sub-factor of *specifications available*. The participant who recommended this factor noted that simply having specifications available is not enough to reverse engineer rendering software for a file format; the available specifications need to be of a high enough quality to do so. Even though this was a qualifying factor for an already discussed factor, participants indicated that it was relevant. This factor received a mean rating value of 1.00, placing it equidistant between somewhat relevant and very relevant. This also placed it among *expertise available* and *legal restrictions* as the third ranked factor for relevancy.

The five participants who rated this factor as *somewhat relevant* stated that higher quality specifications had a direct effect on the endangerment level of a file format. One participant qualified this by writing, “Good quality of the documentation will always be helpful for supporting the use and application of the format.”

The five participants who rated it as *very relevant* also indicated that the quality of the specifications is positively related to the ability to preserve access to the content stored in particular file formats. One participant described this relationship simply, “Good documentation of the format could lead to better preservation efforts.” Another participant noted, “Clear and accurate specifications should reduce the different ways a file format can be implemented and tools built to support it and so the tools for the format should be able to be reimplemented in the same way from a good specification.”

Measurement: Participants who discussed measurement of this factor primarily explained how difficult it would be to measure it. One participant noted, “However it might be difficult to judge what a good specification looks like as the implicit knowledge of the developer and community at that point in time it is written is very difficult to capture - as the fact that ’you don't know what you know‘ is very true here.” Another wrote, “This is very hard to be automatically evaluated. Size of the specifications is not a very good indicator. The size of the textual description per function point could give a clue about the level of completeness for the specification. The quality of the text in terms of understandability level cannot be evaluated effectively in an automatic manner. Probably one indicator could be the ratio of mistakes indicated by a word processor.”

4.3. Questionnaire 4. Special Rater Test and Follow-up Interview

In this section I discuss the results of having an independent, non-expert, special rater search for and apply information on fourteen test endangerment factors to rating the 43 test file formats, and rating each of the factors for relevancy after applying them to rating the formats. After having the special rater participant perform these activities, I asked him to answer several follow-up interview questions about which of the factors he found to be most useful in assessing and most relevant as causes of file format endangerment.

First I asked him how he went about collecting information for each file format endangerment factor. He responded that he typically went first to the Library of Congress' digitalpreservation.gov website and Wikipedia, then tried searches on Google.com for additional information as needed. I then asked him if there were particular file formats for which it was more difficult to find information. He responded: "Yes, I had comparatively more trouble finding information for these formats: kml, xls 5.0, .wpd, hdf, .mov, .vsd, .wmv, .msg, .wmf, .avi, .psd, .rm. These formats often either were not listed on sites like digitalpreservation.gov and/or were not published standards."

When I asked him if there were particular factors that were difficult to find information for, he indicated that locating definitive information on *legal restrictions* and *complexity* was the most difficult. For my final first round question, I asked the special rater how useful he found the factors in helping him rate the endangerment levels, and if there were factors that he found to be more or less useful than others. He responded, "For the most part, I only found the existence and quality of specifications and the existence of rendering software to be useful indicators of endangerment. My rationale is if you can't view the file and you can't easily find specs, it's endangered." He also indicated that the factor, *technical dependencies,* was not a very useful factor he considered when rating the file formats.

After evaluating these responses against the special rater's factor-rating responses, I asked him two additional questions to address some discrepancies between these two sets of responses. First, I asked him, "You stated that *rendering software available*, *specifications available*, and *specification quality* were the most useful indicators of endangerment. However, you rated *rendering software available*, *specifications available*, *ubiquity*, and *community/3rd party support* as *very relevant*, and *specification quality* as *somewhat relevant*. Can you explain the discrepancy between your statement and these ratings?"

He responded first that he rated *specification quality* as *somewhat relevant* because he viewed it as a secondary factor to *specifications available*. He also stated that he viewed *ubiquity* and *community/3rd party support* as secondary factors. Furthermore, he stated, "It was tempting for me to say that ubiquity is a primary indicator, since many formats that are very ubiquitous are not very endangered, but there are also formats that are not widely distributed that are not endangered at all."

I also asked him to name more specifically which sources of information he found to be most useful when answering the questions in the questionnaire. He cited the following online sources: www.digitalpreservation.gov, www.wikipedia.org, and fileformats.archiveteam.org. I made use of the special rater's individual format and factor rating responses in comparison with the Delphi rating and justification text coding discussed in the Results Comparison section, 4.4, below.

4.4. Results Comparison

In this section I compare the results collected from the format rating Delphi, the factor rating Delphi, and the special rater questionnaire and email interview. Through comparing and contrasting the different aspects of these datasets, I triangulate the overall assessment of file format endangerment levels of 43 formats, and of which factors are most relevant as causes/formative indicators of file format endangerment.

I present and discuss three tables containing comparisons of quantitative data collected using the four questionnaires in this study. The first table, *Table 4.4.1*, contains a comparison of the file format ratings collected from the special rater against the mean file format rating means collected from the Delphi participants. The second table, *Table 4.4.2*, contains ranked file format rating data collected from the Questionnaire 2, Round 2 Delphi study; and ranked file format rating data collected from the special rater using Questionnaire 4. The third table, *Table 4.4.3*, contains a comparative set of data collected from three sources: 1) factor appearance count from the file format rating Delphi justification text, 2) mean factor ratings from Questionnaire 3, Round 2 Delphi study, and 3) factor ratings from Questionnaire 4.

The overall mean ratings of the formats were only slightly divergent across datasets. See *Table 4.4.1* for details. The overall Delphi mean score was 1.34 and the overall special rater mean score was 1.14. It is worth noting that the special rater's overall mean score is less than the Delphi mean score; meaning that the special rater score indicates a slightly lower endangerment level. The lower score may be attributed to the fact that the special rater started with data collected by the Delphi participants, and supplemented this baseline data with additional data collection.

Format	Delphi Mean	Special Rater Score	Diff.
.ppt, PowerPoint Presentation	2.33	1.00	-1.33
.rm, Real Media File	2.00	2.00	0.00
.css, Cascading Style Sheet	1.80	2.00	0.20
.wpd, WordPerfect Document	1.78	2.00	0.22
.kmz, Google Earth Placemark File	1.75	1.00	-0.75
.avi, Audio Video Interleave	1.71	2.00	0.29

Format	Delphi Mean	Special Rater Score	Diff.
File			
.psd, Adobe Photoshop Document	1.67	1.00	-0.67
.kml, Keyhole Markup Language	1.67	1.00	-0.67
.xls, Excel Spreadsheet	1.67	2.00	0.33
.wmv, Windows Media Video File	1.63	2.00	0.37
.jpg, Joint Photographic Experts Group File	1.63	1.00	-0.63
.raw, Raw Image Data File	1.58	1.00	-0.58
.hdf, Hierarchical Data Format File	1.57	1.00	-0.57
.doc, Microsoft Word Document	1.56	1.00	-0.56
.vsd, Visio Drawing File	1.50	1.00	-0.50
.php, PHP Source Code File	1.43	1.00	-0.43
.wmf, Windows Metafile	1.43	1.00	-0.43
.rtf, Rich Text Format	1.40	1.00	-0.40
.mov, Apple QuickTime Move	1.38	1.00	-0.38
.pdf, Portable Document Format	1.22	1.00	-0.22
.js, JavaScript File	1.22	1.00	-0.22
.mpg, MPEG Video File	1.22	1.00	-0.22
.docx, Microsoft Word Open XML Document	1.22	1.00	-0.22
.svg , Scalable Vector Graphics	1.22	1.00	-0.22
.xsl, XML Style Sheet	1.20	1.00	-0.20
.pl, Perl Script	1.14	1.00	-0.14

Format	Delphi Mean	Special Rater Score	Diff.
.ico, Icon File	1.13	1.00	-0.13
.c, C Source Code File	1.13	1.00	-0.13
.msg, Microsoft Outlook Email Message	1.11	1.00	-0.11
.pptx, PowerPoint Open XML	1.11	1.00	-0.11
.html, Hypertext Markup Language	1.10	1.00	-0.10
.bmp, Bitmap Image File	1.10	1.00	-0.10
.wav, WAVE Audio File	1.10	1.00	-0.10
.xlsx, Microsoft Excel Open XML Spreadsheet	1.00	1.00	0.00
.for, Fortran Source File	1.00	1.00	0.00
.nc, NetCDF (network Common Data Form)	1.00	1.00	0.00
.xml, Extensible Markup Language	1.00	1.00	0.00
.png, Portable Network Graphic	1.00	1.00	0.00
.txt, Plain Text	1.00	1.00	0.00
.csv, Comma Separated Values	1.00	1.00	0.00
.gif, Graphical Interchange Format	1.00	1.00	0.00
.tif, Tagged Image File Format	1.00	1.00	0.00
.mp3, Moving Picture Expers Group Audio File	1.00	1.00	0.00
Total	**1.34**	**1.14**	**-0.20**

Table 4.4.1. Format rating value differences between the format rating Delphi mean (where -1.00 is '20+ years', 2.00 is '11-20 years', 3.00 is '6-10 years', 4.00 is '1-5 years', and 5.00 is 'already inaccessible') and the special rater score.

Table 4.4.2 shows a comparison of Delphi format rater and special rater ranking of file formats by endangerment level. Though not a direct comparison, since the Delphi mean values are continuous and special rater values are categorical, it provides a relative perspective of which formats each rating group considered to be more endangered than others. For the most part, they are very similar. All six of the formats the special rater rated at 2.00 appear within the top ten most endangered formats rated by the Delphi group. Similarly, the ten file formats that the Delphi group rated at 1.00 were also rated 1.00 by the special rater.

Table 4.4.3 shows a comparison of ranked factors in order of prevalence (in the case of the format rating justification text count) and rating level (Delphi factor rating means and special rater ratings). Examining each dataset included in this table reveals cutoff points for which factors are the most important for indicating file format endangerment. In the Delphi format rating justification text coding count data, there is a distinct drop-off of factor appearances after *specifications available*. While *legal restrictions* appeared in the format rating justification text 97 times, the next most frequently appearing factor, *complexity*, only appeared 63 times. This leaves *rendering software available*, *ubiquity*, *specifications available*, and *legal restrictions* as well-agreed-upon factors to consider in further analysis.

A logical cutoff point for both the Delphi factor rating mean ranking and special rater factor ratings datasets is a rating above 1.00, the halfway point between *somewhat relevant* and *very relevant*. A rating above 1.00 indicates that the factor was rated close to *very relevant*, whereas factors rated at or below 1.00 are at most *relevant*. For the Delphi factor rating mean ranking this leaves the factors *specifications available*, *rendering software available*, *technical dependencies*, and *community/3rd party support*. For the special rater factor ratings this leaves *rendering software available*, *specifications available*, *ubiquity*, and *community/3rd party support*.

Delphi Format Rating Mean Ranking	Special Rater Format Ratings
.ppt, PowerPoint Presentation (2.33)	**.wmv** Windows Media Video File (2.00)
.rm, Real Media File (2.00)	**.rm** Real Media File (2.00)
.css, Cascading Style Sheet (1.80)	**.css** Cascading Style Sheet (2.00)
.wpd, WordPerfect Document (1.78)	**.wpd** WordPerfect Document (2.00)
.kmz, Google Earth Placemark File (1.75)	**.xls** Excel Spreadsheet (2.00)
.avi, Audio Video Interleave File (1.71)	**.avi** Audio Video Interleave File (2.00)
.psd, Adobe Photoshop Document (1.67)	**.kml** Keyhole Markup Language (1.00)
.kml, Keyhole Markup Language (1.67)	**.kmz** Google Earth Placemark File (1.00)
.xls, Excel Spreadsheet (1.67)	**.raw** Raw Image Data File (1.00)
.wmv, Windows Media Video File (1.63)	**.psd** Adobe Photoshop Document (1.00)
.jpg, Joint Photographic Experts Group File (1.63)	**.jpg** Joint Photographic Experts Group File (1.00)
.raw, Raw Image Data File (1.58)	**.ppt** PowerPoint Presentation (1.00)
.hdf, Hierarchical Data Format File (1.57)	**.hdf** Hierarchical Data Format File (1.00)
.doc, Microsoft Word Document (1.56)	**.doc** Microsoft Word Document (1.00)

Delphi Format Rating Mean Ranking	Special Rater Format Ratings
.vsd, Visio Drawing File (1.50)	**.vsd** Visio Drawing File (1.00)
.php, PHP Source Code File (1.43)	**.php** PHP Source Code File (1.00)
.wmf, Windows Metafile (1.43)	**.wmf** Windows Metafile (1.00)
.rtf, Rich Text Format (1.40)	**.rtf** Rich Text Format (1.00)
.mov, Apple QuickTime Move (1.38)	**.mov** Apple QuickTime Move (1.00)
.pdf, Portable Document Format (1.22)	**.pdf** Portable Document Format (1.00)
.js, JavaScript File (1.22)	**.js** JavaScript File (1.00)
.mpg, MPEG Video File (1.22)	**.mpg** MPEG Video (1.00)
.docx, Microsoft Word Open XML Document (1.22)	**.docx** Microsoft Word Open XML Document (1.00)
.svg , Scalable Vector Graphics (1.22)	**.svg** Scalable Vector Graphics (1.00)
.xsl, XML Style Sheet (1.20)	**.xsl** XML Style Sheet (1.00)
.pl, Perl Script (1.14)	**.pl** Perl Script (1.00)
.ico, Icon File (1.13)	**.ico** Icon File (1.00)
.c, C Source Code File (1.13)	**.c** C Source Code File (1.00)
.msg, Microsoft Outlook Email Message (1.11)	**.msg** Microsoft Outlook Email Message (1.00)
.pptx, PowerPoint Open XML (1.11)	**.pptx** PowerPoint Open XML (1.00)

Delphi Format Rating Mean Ranking	Special Rater Format Ratings
.html, Hypertext Markup Language (1.10)	**.html** Hypertext Markup Language (1.00)
.bmp, Bitmap Image File (1.10)	**.bmp** Bitmap Image (1.00)
.wav, WAVE Audio File (1.10)	**.wav** WAVE Audio File (1.00)
.xlsx, Microsoft Excel Open XML Spreadsheet (1.00)	**.xlsx** Microsoft Excel Open XML Spreadsheet (1.00)
.for, Fortran Source File (1.00)	**.for** Fortran Source File (1.00)
.nc, NetCDF (network Common Data Form) (1.00)	**.nc** NetCDF (network Common Data Form) (1.00)
.xml, Extensible Markup Language (1.00)	**.xml** Extensible Markup Language (1.00)
.png, Portable Network Graphic (1.00)	**.png** Portable Network Graphic (1.00)
.txt, Plain Text (1.00)	**.txt** Plain Text (1.00)
.csv, Comma Separated Values (1.00)	**.csv** Comma Separated Values (1.00)
.gif, Graphical Interchange Format (1.00)	**.gif** Graphical Interchange Format (1.00)
.tif, Tagged Image File Format (1.00)	**.tif** Tagged Image File Format (1.00)
.mp3, Moving Picture Expers Group Audio File (1.00)	**.mp3** Moving Picture Expers Group Audio File (1.00)

Table 4.4.2. Ranked format rating comparison chart.

Delphi Format Rating Justification Text Coding Count	Delphi Factor Rating Mean Ranking	Special Rater Factor Ratings
Rendering Software Available (162)	Specifications Available (1.40)	Rendering Software Available (1.50)
Ubiquity (130)	Rendering Software Available (1.10)	Specifications Available (1.50)
Specifications Available (111)	Technical Dependencies (1.10)	Ubiquity (1.50)
Legal Restrictions (97)	Community/3rd Party Support (1.10)	Community/3rd Party Support (1.50)
Complexity (63)	Expertise Available (1.00)	Legal Restrictions (0.50)
Community/3rd Party Support (51)	Legal Restrictions (1.00)	Technical Dependencies (0.50)
Specification Quality (46)	Specification Quality (1.00)	Standardization (0.50)
Developer/Corporate Support (44)	Ubiquity (0.90)	Specification Quality (0.50)
Standardization (42)	Cost (0.80)	Backward/Forward Compatibility (0.50)
Technical Dependencies (42)	Complexity (0.70)	Ease of Identification (0.50)
Rendering Software Feature/Functionality/Behavior Support (18)	Standardization (0.60)	Expertise Available (0.50)
Backward/Forward Compatibility (12)	Backward/Forward Compatibility (0.60)	Cost (-0.50)
Value (11)	Developer/Corporate Support	Complexity (-0.50)

Delphi Format Rating Justification Text Coding Count	Delphi Factor Rating Mean Ranking	Special Rater Factor Ratings
	(0.60)	
Compression (10)	Ease of Identification (0.50)	Developer/Corporate Support (-0.50)
Lifetime (8)	Lifetime (0.40)	Ease of Validation --
Ease of Identification (3)	Availability Online (0.40)	Value --
Technical Protection Mechanism (1)	Domain Specificity (0.30)	Compression --
Domain Specificity (1)	Technical Protection Mechanism (0.30)	Lifetime --
Cost --	Value (0.30)	Technical Protection Mechanism --
Revision Rate --	Revision Rate (0.25)	Revision Rate --
Institutional Policies --	Geographic Spread (0.20)	Institutional Policies --
Ease of Validation --	Ease of Validation (0.20)	Rendering Software Feature/Functionality/ Behavior Support --
Geographic Spread --	Institutional Policies (0.20)	Geographic Spread --
Error-tolerance --	Compression (-0.10)	Error-tolerance --
Metadata Support --	Error-tolerance (-0.20)	Metadata Support --
Storage Space --	Metadata Support (-0.20)	Storage Space --

Delphi Format Rating Justification Text Coding Count	Delphi Factor Rating Mean Ranking	Special Rater Factor Ratings
Viruses --	Storage Space (-0.20)	Viruses --
Availability Online --	Viruses (-0.40)	Availability Online --
Expertise Available --	Rendering Software Feature/Functionality/Behavior Support --	Domain Specificity --

Table 4.4.3. Ranked factor comparison chart with delineated cutoff points.

CHAPTER 5: DISCUSSION

Each set of data collected through the three studies presented here provided valuable insights into the nature of file format endangerment. In this chapter, I discuss the results of the studies individually, and then I compare the datasets with each other and explore the implications of the body of data as a whole. I discuss the implications of the findings for my four research questions and for continued research in this area.

5.1. Questionnaire 2. File Format Rating

Two important things resulted from the format rating Delphi activities. One was a greater understanding of endangerment levels of the fifty file formats, and the other was a greater understanding of which factors expert participants considered when rating file format endangerment levels.

Overall, the expert participants rated the file formats to be not very endangered. This finding stands in stark contrast to the amount of literature that discusses file formats and the risks they pose to digital preservation. Of the forty-three file formats rated in the second format rating Delphi round, only two formats had a mean rating above 2.00 (Power Point 2.0 for Windows and Real Media 4.0.1), where 2.00 indicated the choice, “Information stored in this file format will be inaccessible in 11-20 years,” and was the second lowest rating that was available. Of these two, the highest rating was 2.33 for Power Point 2.0 for Windows. Twenty-nine of the forty-three formats (67%) were rated at or below 1.50, ten of which (23%

of the total) were rated at 1.00, the lowest possible rating.

This finding aligns with David Rosenthal's (2010) assertions that digital content is not at risk of becoming inaccessible due to the format in which it is encoded. It is also important to consider the fact that participants in this part of the study reported a high level of technical experience with managing file formats in a digital preservation context. The qualitative data collected during this Delphi study reveals that these participants believe that information can be accessed from most of the file formats because they have the technical expertise to do it themselves. This level of expertise and experience accessing information from a wide variety of file formats may also have influence over some participants' lower ratings of how important they believed file formats were to continued access to digital content. If they had the technical expertise to access content stored in most file formats, they likely do not see file formats as a large threat to sustaining access to digital content.

Participants in this study have demonstrated experience in this area and mostly work in larger institutions that have a greater capacity to support solutions for accessing content encoded in more challenging file formats. In the wider world, many professionals working in memory institutions probably do not have the same level of expertise, and most importantly do not work for institutions with the same capacity to support complex digital preservation research. Based on this, it can further be assumed that if working professionals were to have participated in this study, they would have rated the file formats at higher endangerment levels. By extension, if these working professionals had access to the expertise of participants in this study and/or the same level of institutional support, they might rate the file formats at a lower endangerment level. This reasoning highlights the need to connect the expertise exhibited by participants in this study with those who are not aware of some of the methods and tools that can be used to access information stored in apparent high-risk file formats.

The second set of results emerged through coding the participant comment text. In examining the comment text, I discovered seventeen themes that participants cited as factors explaining their file format endangerment level ratings: *rendering software available, ubiquity, specifications available, legal restrictions, complexity, community/3rd party support, specification quality, developer/corporate support, standardization, technical dependencies, backward/forward compatibility, rendering software/feature/functionality/behavior support, value, compression, lifetime, ease of identification,* and *technical protection mechanism*.

The top four factors mentioned were *rendering software available* (162 mentions), *ubiquity* (130 mentions), *specifications available* (111 mentions), and *legal restrictions* (97 mentions). The factor with the next highest number of mentions, *complexity*, had only 63 mentions. There is a 42.5% change in the number of mentions of *legal restrictions* and *complexity*, whereas the top four factors had much smaller changes: *ubiquity* had 21.9% fewer mentions than *rendering software available; specifications available* had 15.8% fewer mentions than *ubiquity*; *legal restrictions* had 13.5% fewer mentions than *specifications available*. This sharp drop in mentions between *legal restrictions* and *complexity* suggests a reasonable cutoff point for further investigation.

It is important to recognize the implications associated with the seven formats that were removed from the study because too few participants had enough experience with them to rate the formats well. Lack of knowledge about and experience with file formats, especially when it pertains to finding or creating the software necessary to render it, can greatly impact a format's rated endangerment level. Though these formats were not rated a second time in the Delphi study, their first round endangerment levels were relatively high. Three of the seven removed formats (43%) were rated above 2.00, or "11-20 years," while only two formats of the remaining 43 (5%) were rated above 2.00 in both Round 1 and Round 2.

A general lack of information and knowledge about particular file formats (even among participants recruited for their expertise in file formats) can be a strong indicator that the community should undertake efforts to discover and document the information necessary to maintain access to contents encoded in those formats. Furthermore, this could be an indicator that there is possibly not enough information about these file formats to maintain access to the contents encoded within them. Using the file format endangerment index to guide data collection and endangerment level ratings on these formats could reveal that content encoded in these file formats are at risk of becoming inaccessible, which could trigger actions to address the areas that influence this risk.

5.2. Questionnaire 3. Factor Rating

I asked expert participants to rate factors for relevancy as a cause of file format endangerment in order to make sense of the dozens of factors discussed in the literature and to elicit their views on which of the factors have a direct effect on the ability to access information encoded within a particular file format. Both the numerical ratings and participant comments provided insight into this issue.

First, the numerical ratings provided a cutoff for which factors participants believed were at least *somewhat relevant*. With the *somewhat relevant* rating having a value of 0.50, anything that received a rating below 0.50 did not make the cutoff. Half of the factors were rated at 0.50 and above. This cutoff allowed me to eliminate the half of the factors that were rated below 0.50, focusing instead on those factors that the experts deemed to be most relevant. No factor received unanimous ratings of *very relevant*.

Only six factors were rated at 1.00, which is the halfway point between *somewhat relevant* and *very relevant*. If I were selecting factors based solely on the data collected from this Delphi study, this would be the most logical cutoff point, as 1.00 is a good candidate value for a simply "relevant" rating. The factors that were rated at 1.00 and above were: *specification quality* (1.00), *expertise available* (1.05), *community/3rd party support* (1.05), *technical dependencies* (1.05), *rendering software available* (1.14), and s*pecifications available* (1.41).

The comments from participants provided insight into the complex nature of the issue. Many of the comments reflected the ambiguity of some of the factors. For example, one participant wrote about *complexity*, "This is an 'it depends' answer - complexity is hard to bundle into one type of characteristic. Different types of complexity could be answered on their own." Another wrote on the *cost* factor, "I agree with round 1 responses that state cost as a complex, multi-faceted and organizational[ly] influenced factor." Other factors proved to be less ambiguous and participants were able to more directly justify their ratings.

The fact that only six factors were rated at 1.00 and above is an important finding. I began this research with a total of 138 individual factors that I found in the literature. I was able to reduce this list of factors to 21 factors. Through the Delphi process, I was then able to reduce this number to six factors that participants rated as at least halfway between *somewhat relevant* and *very relevant*. Reducing the number of factors this amount was a large step toward the final selection of clear formative indicators for a file format endangerment index.

5.3. Questionnaire 4. Special Rater Test and Follow-up Interview

The purposes of the special rater activities were to compare the special rater format rating results with the Delphi format rating results, and to test the usefulness of the factors by directly applying them to evaluating file format endangerment for each of the test formats. By asking the special rater to collect information on each of the fourteen factors presented, the special rater was able to experience first-hand the factors for which useful information is available and which factors informed a realistic endangerment rating for the format. The primary intention was to test the application of the factors.

After collecting information for each of the fourteen factors that were selected as a result of the factor-rating Delphi, the special rater rated the file formats using the same scale as the Delphi format raters. Of the forty-three formats rated, the special rater only rated six formats (14%) at 2.00, or 11-20 years. The special rater rated the remaining 86% of the formats at the lowest endangerment level of 1.00, or 20+ years. The overall mean score of these ratings was very low, at 1.14, which suggests a lower risk associated with file formats than has generally been indicated in the literature.

This phase of the study yielded lessons about which factors proved most helpful in rating file format endangerment. According to the participant's post hoc factor ratings, only four were rated as *very relevant*: *rendering software available*, *specifications available*, *ubiquity*, and *community/3rd party support*. However, in the follow up interview the participant stated, "For the most part, I only found the existence and quality of specifications and the existence of rendering software to be useful indicators of endangerment. My rationale is if you can't view the file and you can't easily find specs, it's endangered." The participant's file format rating justification comments, however, contains references to a number of other factors. For example, the participant mentions *standardization*, *backward/forward compatibility*, *complexity*, and *developer/corporate support* in several comments.

When questioned further about these apparent discrepancies, the participant answered, “I see ‘ubiquity’ and ‘community/3rd party support’ as really secondary indicators. It was tempting for me to say that ubiquity is a primary indicator, since many formats that are very ubiquitous are not very endangered, but there are also formats that are not widely distributed that are not endangered at all, such as the .nes format, used for ROM dumps of Nintendo Entertainment System cartridges.” In summary, the special rater asserted that rendering *software available* and *specifications available* were the most useful factors in rating file format endangerment levels.

5.4. Results Comparison

After comparing the results from the three sets of collected data, five factors emerged as being either more highly ranked, or as appearing more times in the format-rating justification text. Examining each of the five remaining factors in light of the qualitative data collected provides more clarity for which are the most relevant as candidate causes of file format endangerment.

Rendering software available. *Rendering software available* and *specifications available* are the only two factors that appeared beyond the cutoff point in all three datasets. It appeared as the top factor in two of the three datasets, and would have tied for the top ranking in the Delphi factor rating dataset if not for one *not relevant at all* rating. The rationale for this aberrant rating was justified that the participant considered the lack of rendering software to be the definition of obsolescence/file format endangerment and therefore rated it as being not relevant within the context of the participant's self-selected definition.

Four of the eight participants who rated this factor as *very relevant* indicated lack of rendering software strongly suggests file format obsolescence. For example, one participant wrote, "By definition without rendering software the format is obsolete." By far, the comments about the *rendering software* factor in the Delphi factor rating exercise were very strong, simple, and direct: without rendering software a file format is essentially obsolete. The strength of the comments about this factor points to it being a very strong candidate as a direct cause of file format endangerment.

Specifications available. Like *rendering software available*, the *specifications available* factor was included beyond the cutoff point in all three factor evaluation datasets in this study. It received a very high relevancy rating (1.40 of 1.50 possible) from the Delphi factor rating participants. Delphi participants indicated that having specifications available enables the creation of rendering software if none is available. Furthermore, others indicated that it helps to determine if software faithfully renders the contents of a file. One participant wrote, "It is hard to see that a format would not be more endangered if specifications could not be obtained." Based on the ratings and the strength of the participant comments, the *specifications available* factor is another strong candidate as a cause of file format endangerment.

Ubiquity. The case for considering the *ubiquity* factor as a cause of file format endangerment is weakened for several reasons. First is the fact that it only remained above the cutoff point in two of the three datasets. Second, though the special rater rated it as *very relevant*, he explained later that he only considered it to be a secondary factor, because of the following scenario: “there are also formats that are not widely distributed that are not endangered at all, such as the .nes format, used for ROM dumps of Nintendo Entertainment System cartridges.”

This sentiment is echoed in many of the Delphi factor rating comments, where several participants described its effect on endangerment in secondary terms. For example, one participant wrote, “The popularity of a given file format increases the support provided by user communities and consequently increases the resources allocated/available for development/maintenance for further developments.” In this scenario, the ubiquity of the file format has an effect on other factors that directly affect the endangerment level of the format and serves more as a tertiary factor that affects *community/3rd party support*.

Community/3rd party support. This factor is ultimately a secondary factor, even though it appeared above the cutoff point in two of the three datasets. Participants in the factor rating Delphi referred to it as a stopgap against a single point of failure: “single-point of failures are serious potential problems, and having a format which is supported by a single provider, rather enjoying larger community and 3rd party support, is a classic single point of failure situation. The wider the experience with and understanding of a format, the better, and the lack of those can present serious risks.” In this case, community/3rd party support is a factor that can directly support the existence of rendering software, but is often contingent on the availability of specifications.

Technical dependencies. This factor appeared above the cutoff line in only the Delphi factor rating dataset. The special rater noted that he “didn't find technical dependencies to be a useful indicator as all formats have some technical dependencies.” When the format rating Delphi participants mentioned technical dependencies, it was typically in the context of causing problems with the full and faithful rendering of a file that calls in information from external files; but do not mention it preventing a file from being rendered at all. In this case, *technical dependencies* is a tertiary factor where *rendering software* is the primary and *rendering software feature/functionality/behavior support* is the secondary factor.

Legal restrictions. This factor appeared above the cutoff line in only the Delphi format justification text coding dataset. Close examination of temporal priority reveals that while legal restrictions do have an effect on accessibility of digital content, this factor is actually a secondary factor to *specifications available* and *community/3rd party support*. The instances where legal restrictions were coded in the format rating justification text were those times where participants mentioned the availability of specifications and the existence of open source software. Legal restrictions can prohibit the free availability of specifications and prohibits the creation of rendering software through third parties.

It was through the process of comparing these results that I was able to make a final reduction in factors from six to three: *rendering software available*, *specifications available*, and *community/3rd party support.* From beginning to end, I was able to reduce the list of factors from the original 138 factors that I found in the literature to three, for a total reduction of 135 factors.

5.5. Test Application of Index Factors and Rating Guide

In this section, I apply the information collected for the three selected factors – *rendering software available*, *specifications available*, and *community/3rd party support* – by the special rater in Questionnaire 4 to a simple file format evaluation-scoring test. For each file format, a score will be added in increments of 0.50 for each time it is indicated that 1) no rendering software is available, 2) no specifications are available, and 3) it is indicated that there is little to no community/3rd party support. A low endangerment score would be zero and a high endangerment score would be 1.50.

Format	Rendering Software Available	Specifications Available	Community/3rd Party Support	Score	Mean
.nc NetCDF (network Common Data Form)	Software available in Java, C, and Fortran versions: http://www.unidata.ucar.edu/downloads/netcdf/index.jsp **Score: 0**	Specifications available here: http://www.opengeospatial.org/standards/netcdf More structure here: http://www.unidata.ucar.edu/software/netcdf/docs /faq.html#format **Score: 0**	Published standard at Open Geospatial Consortium: http://www.opengeospatial.org/standards/netcdf Used at research institutions across the world: http://www.unidata.ucar.edu/software/netcdf/usage.html **Score: 0**	0.00	1.00
.xml Extensible Markup Language	Many pieces of software can render xml. **Score: 0**	http://www.w3.org/TR/REC-xml/ **Score: 0**	Widely used in a variety of contexts. **Score: 0**	0.00	1.00
.pdf Portable Document Format	Widespread **Score: 0**	1.0 specs not readily available **Score: 0.50**	Many other developers use their own or open source libraries to render newer versions of pdf files. **Score: 0**	0.50	1.00

.png Portable Network Graphic	Many: http://www.libpng.org/pub/png/pngapps.html **Score: 0**	http://www.libpng.org/pub/png/pngdocs.html **Score: 0**	In wide use by 3rd parties: http://www.libpng.org/pub/png/pngapps.html **Score: 0**	0.00	1.00
.kml Keyhole Markup Language	Many GIS apps listed at http://www.digitalpreservation.gov/formats/fdd/fdd000340.shtml **Score: 0**	http://www.opengeospatial.org/standards/kml **Score: 0**	Moderate, some software out there that uses it. **Score: 0**	0.00	1.67
.txt Plain Text	Just about everything that can read any sort of document. **Score: 0**	Not Applicable **Score: 0**	In use in every piece of software ever released, more or less. **Score: 0**	0.00	1.00
.csv Comma Separated Values	Widespread availability **Score: 0**	http://tools.ietf.org/html/rfc4180 **Score: 0**	Widely supported by 3rd party applications, easy to code your own interpreter with any scripting language **Score: 0**	0.00	1.00
.gif Graphical Interchange Format	Many **Score: 0**	http://electronic-records-preservation.ourarchives.wikispaces.net/file/view/GIF87a.pdf http://www.fileformat.info/format/gif/spec/index.htm **Score: 0**	High. In use by just about any program that can read graphics **Score: 0**	0.00	1.00

.html Hypertext Markup Language	Too many to count **Score: 0**	http://tools.ietf.org/html/rfc1866 **Score: 0**	Wide **Score: 0**	0.00	1.10
.jpg Joint Photographic Experts Group	Yes, too numerous to list **Score: 0**	http://www.jpeg.org/jpeg/ **Score: 0**	Wide **Score: 0**	0.00	1.63
.xls Excel Spreadsheet	MS Office, other office suites **Score: 0**	MS "Open Specification Promise" http://www.digitalpreservation.gov/formats/intro/specifications.shtml does not include version 5 **Score: 0**	Can be rendered with varying degrees of success by open source office suites like LibreOffice and OpenOffice **Score: 0**	0.00	1.67
.tif Tagged Image File Format	Widespread **Score: 0**	yes: http://www.digitalpreservation.gov/formats/fdd/fdd000022.shtml#specs **Score: 0**	Widely supported in open source community through libtiff library **Score: 0**	0.00	1.00
.wpd WordPerfect Document	Open source office suites **Score: 0**	no longer easily findable **Score: 0.50**	Renderable by libwpd, notes on how to extract text on http://justsolve.archiveteam.org/wiki/WordPerfect **Score: 0**	0.50	1.78

.doc Microsoft Word Document	MS Office, open source office suites, MS Word Viewer **Score: 0**	yes: http://www.digitalpreservation.gov/formats/digformatspecs/Word97-2007BinaryFileFormat%28doc%29Specification.pdf **Score: 0**	Reverse engineered by OpenOffice per http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=690 **Score: 0**	0.00	1.56
.hdf Hierarchical Data Format File	Yes http://www.hdfgroup.org/downloads/index.html **Score: 0**	yes http://www.hdfgroup.org/release4/doc/DSpec_html/DS. **Score: 0**	Reasonably large niche community: http://www.hdfgroup.org/users.html **Score: 0**	0.00	1.57
.kmz Google Earth Placemark File	Google Earth, Google Maps, 3D Route Builder per http://file.org/extension/kmz# **Score: 0**	https://developers.google.com/kml/documentation/kmzarchives?csw=1%20ar **Score: 0**	Some other geo software packages support kmz beyond Google, see short list http://file.org/extension/kmz **Score: 0**	0.00	1.75
.mp3 Moving Picture Expers Group Audio File	Many **Score: 0**	ISO/IEC 11172-3:1993. Information technology -- **Score:**	Widely used outside of Motion Pictures Expert **Score: 0** Group, original developers. Open source coder/decoder software widely available	0.00	1.00

.ppt PowerPoint Presentation	MS and open source office suites **Score: 0**	no, only for newer versions **Score: 0.50**	Some open source office suites may be able to render this version with unknown fidelity **Score: 0**	0.50	2.33
.mov Apple QuickTime Move	Many, both from Apple and open source like VLC. **Score: 0**	Yes: https://developer.apple.com/standards/qtff-2001.pdf **Score: 0**	Renderable widely: http://www.fileinfo.com/extension/mov **Score: 0**	0.00	1.38
.c C Source Code File	Many open source compilers **Score: 0**	Yes **Score: 0**	Wide. Many operating systems are written in C or flavors of C **Score: 0**	0.00	1.13
.vsd Visio Drawing File	http://libregraphicsworld.org/blog/entry/initial-support-for-visio-files-lands-to-libreoffice and MS Office **Score: 0**	In as much as any specs exist: http://libregraphicsworld.org/blog/entry/re-lab-reverse-engineers-visio-file-formats-publishes-the-first-spec **Score: 0**	Format reverse engineered: http://libregraphicsworld.org/blog/entry/re-lab-reverse-engineers-visio-file-formats-publishes-the-first-spec; talk of integrating it into LibreOffice **Score: 0**	0.00	1.50
.js JavaScript File	Yes, open source renderers available **Score: 0**	ECMAScript v 3: http://www.ecmascript.org/docs.php **Score: 0**	Widely used in both client and server side web work. **Score: 0**	0.00	1.22

.css Cascading Style Sheet	Any contemporary web browser, assuming content to be styled also present **Score: 0**	http://www.w3.org/TR/CSS21/ **Score: 0**	CSS rendering engines present in all major web browser engines **Score: 0**	0.00	1.80
.xsl XML Style Sheet	Yes, open source libraries **Score: 0**	yes, W3C spec **Score: 0**	A number of open source libraries support xsl **Score: 0**	0.00	1.20
.raw Raw Image Data File	many open source packages can view tiffs **Score: 0**	yes, ISO 12234-2:2001 standard **Score: 0**	Used by some cameras as a raw image format, like Nikon **Score: 0**	0.00	1.57
.rtf Rich Text Format	Many **Score: 0**	yes: http://msdn.microsoft.com/en-us/library/office/aa140277(v=office.10).aspx **Score: 0**	Open source office suites support rtf **Score: 0**	0.00	1.40
.bmp Bitmap Image File	Many **Score: 0**	yes, see http://www.digitalpreservation.gov/formats/fdd/fdd000189.shtml **Score: 0**	Widely renderable by 3rd party viewers b/c of simplicity of format per http://www.digitalpreservation.gov/formats/fdd/fdd000189.shtml **Score: 0**	0.00	1.10
.mpg MPEG Video File	Many open source renderers **Score: 0**	yes: http://www.digitalpreservation.gov/formats/fdd/fdd000035.shtml#specs **Score: 0**	Widely used by 3rd party software and by GLAM institutions for archiving **Score: 0**	0.00	1.22

.docx Microsoft Word Open XML Document	MS and other office suites **Score: 0**	Yes **Score: 0**	Can be opened by some open source office suites **Score: 0**	0.00	1.22
.wmv Windows Media Video File	Proprietary and open source renderers available **Score: 0**	Given open source renderers, it's gotta be somewhere! not easily findable **Score: 0**	Can be played by VLC and some other community players **Score: 0**	0.00	1.63
.wav WAVE Audio File	Many **Score: 0**	yes: http://www.digitalpreservation.gov/formats/fdd/fdd000001.shtml **Score: 0**	Widely used in audio programs **Score: 0**	0.00	1.10
.php PHP Source Code File - Hypertext Preporcessor	php can render php, that may not be all you need to render the output of a php script **Score: 0**	yes, largely in code **Score: 0**	widely used **Score: 0**	0.00	1.43
.msg Microsoft Outlook Email Message	Outlook **Score: 0**	yes: http://msdn.microsoft.com/en-us/library/cc463912(v=exchg.80).aspx **Score: 0**	Moderate. Details about how to read files are here: http://www.fileformat.info/format/outlookmsg/ **Score: 0**	0.00	1.11
.svg Scalable Vector Graphics	Many graphics packages support libsvg open source library **Score: 0**	http://www.w3.org/TR/SVG11/ **Score: 0**	Yes, renderable by many graphics packages **Score: 0**	0.00	1.22

.wmf Windows Metafile	Yes, ACDSee, IrfanView, and MS Office applications **Score: 0**	http://msdn.microsoft.com/en-us/library/cc215212.aspx **Score: 0**	Some viewers can open it, like ACDSee and IrfanView **Score: 0**	0.00	1.43
.AVI Audio Video Interleave File	VLC, other players **Score: 0**	http://msdn.microsoft.com/en-us/library/windows/desktop/dd318187(v=vs.85).aspx **Score: 0**	Can be viewed by VLC **Score: 0**	0.00	1.71
.psd Adobe Photoshop Document	Photoshop, GIMP **Score: 0**	http://www.adobe.com/devnet-apps/photoshop/fileformatashtml/#50577409_19840 **Score: 0**	GIMP partially supports PSD **Score: 0**	0.00	1.67
.for Fortran Source File	Fortran compilers **Score: 0**	NIST published FIPS PUB 69, which specifics Fortran 77 syntax per http://en.wikipedia.org/wiki/FORTRAN_77#FORTRAN_77 **Score: 0**	Still in use in scientific computing community **Score: 0**	0.00	1.00
.pptx PowerPoint Open XML	MS Office, open source suites **Score: 0**	http://www.digitalpreservation.gov/formats/digformatspecs/PowerPoint97-2007BinaryFileFormat%28ppt%29Specification.pdf **Score: 0**	Supported partially by open source office suites **Score: 0**	0.00	1.11

.rm Real Media File	RealPlayer Cloud, Real Alternative **Score: 0**	https://common.helixcommunity.org/2003/HCS_SDK_r5/htmfiles/rmff.htm **Score: 0**	Some, Real Alternative player can play it back: http://en.wikipedia.org/wiki/Real_Alternative#Real_Alternative **Score: 0**	0.00	2.00
.xlsx Microsoft Excel Open XML Spreadsheet	Office, Excel Viewer, OpenOffice **Score: 0**	http://msdn.microsoft.com/en-us/library/dd922181%28v=office.12%29.aspx **Score: 0**	Readable in OpenOffice per http://www.openoffice.org/dev_docs/features/3.0/ **Score: 0**	0.00	1.00
.pl Perl Script	Perl compiles on many operating systems **Score: 0**	perl.org only publicly shows docs back through 5.8; they may still have 5.6 docs **Score: 0**	Widely used on Unix-like operating systems **Score: 0**	0.00	1.15
.ICO Icon File	Yes, any browser can render .icos, and many viewers like IrfanView and ACDSee can view ICO files **Score: 0**	http://msdn.microsoft.com/en-us/library/ms997538.aspx **Score: 0**	Used to deliver website icons by apache and other open source web servers, readable by many open source viewers **Score: 0**	0.00	1.13

Table 5.5.1. Special rater factor data with simple scoring test.

In applying this simple test, only three file formats received a score above 0.00: Portable Document Format 1.0, WordPerfect Document 6.2, and PowerPoint Presentation 2.0 for Windows, which all received a score of 0.50 and both due to a lack of available specifications. Two of these formats, WordPerfect Document 6.2 and PowerPoint Presentation 2.0 for Windows, received high endangerment ratings during the format-rating

Delphi process. PowerPoint Presentation 2.0 for Windows received the highest mean score of 2.33 during the Delphi rating exercise, and WordPerfect Document 6.2 received the fourth highest mean score at 1.78. The fact that both of these formats received high mean ratings and high test endangerment scores indicates a potentially strong relationship between the three selected factors and a file format's endangerment level.

However, the format Portable Document Format 1.0 received a relatively low mean score of 1.22 in the Delphi rating process. It is unclear why this format received a low mean score, though it is possible that participants were rating it as a general format and not the specific version listed. Examining the justification text from the Round 2 format-rating Delphi exercise shows that only three participants discussed the version of the format in their rating justifications. Two of these participants were the only two who rated it at 11-20 years, where the remaining participants rated it at 20+ years. Examining the justification text of the remaining seven participants reveals that they were not considering the version of the format as they rated the format. This is most likely the reason for the lower mean rating.

5.5.1. File Format Endangerment Rating Guide

Assuming that the three factors applied above – rendering software available, specifications available, and community/3rd party support – are tested and validated as appropriate formative indicators for a file format endangerment index, they can be applied as a guide for file format assessments in the field. An individual may collect information on these three factors and apply a similar rating formula to determine if a particular file format presents risk to the continued access of the contents encoded in it. In this case, any score above 0.00 would represent enough risk that the format may be considered endangered, and therefore actions should be taken to either migrate the content away from that format, to recreate viable rendering software for the format, or to create a viable emulation platform from which the contents may be accessed.

5.6. Implications

In reviewing and comparing the collected data, several findings were apparent in relation to my four research questions, discussed in detail here. I also discuss additional implications for applications of this research.

RQ1: *Do digital preservation experts believe that certain file formats pose a risk for digital preservation?*

Overall, the mean file format ratings indicate that experts believe that most of the file formats they rated were not particularly endangered. Of the 43 formats that were not removed from the Delphi study after Round 1, only two were rated at or above 2.00, which indicates the rating choice "information encoded in this file format will be inaccessible in 11-20 years." No file formats had mean ratings at 3.00 (inaccessible in 6-10 years), 4.00 (1-5 years), or 5.00 (already inaccessible).

Based on this data, it is clear that the expert participants did not believe that certain file formats pose an immediate risk for digital preservation. According to the mean ratings, participants as a whole believe that the soonest content encoded in a particular file format will become inaccessible is in 11-20 years.

RQ2: *Which file formats do digital preservation experts believe are more endangered than others?*

The two file formats rated above 2.00 were PowerPoint Presentation, Version 2.0 for Windows (2.33) and Real Media File, Version 4.01 (2.00). Of the seven formats removed after Round 1 of the Delphi process three were rated above 2.00: Microsoft Access Database, Version 7.0 for Windows (2.25), Lotus 1-2-3 Worksheet, Version 2.0 (2.20), and StarOffice Writer Text Document, Version 5.0 (2.20).

RQ3: *What are the most relevant formative indicators of file format endangerment, and how can these indicators be measured?*

Considering all of the findings, the only factor that can be considered a direct cause to file format endangerment is *rendering software available*. Secondary factors to this are *specifications available* and *community/3rd party support*. *Ubiquity* and *technical dependencies* are tertiary factors that would not add sufficient understanding to file format endangerment levels. As a primary factor, *rendering software available* should be included in a file format endangerment index. The secondary factors, *specifications available* and *community/3rd party support* are useful in cases where there is no rendering software available, as the ability to create rendering software where none exists is often contingent on the existence of one or both of these factors.

RQ4: *How effectively can the expert-chosen file format endangerment factors be applied to rating file format endangerment?*

Of the 14 factors the factor rating Delphi experts rated above 0.50 (between *somewhat relevant* and *very relevant*), the Questionnaire 4 special rater rated four as *very relevant*: *rendering software available*, *specifications available*, *ubiquity*, and *community/3rd party support*. In the follow-up interview, the special rater indicated that only two of those four are actually relevant: *rendering software available*, *specifications available.*

The special rater noted that the factors *legal restrictions* and *complexity* were more difficult than others to find information for online. Additionally, he indicated that there were some file formats about which it was more difficult to find factor information. These formats were: kml, xls 5.0, .wpd, hdf, .mov, .vsd, .wmv, .msg, .wmf, .avi, .psd, .rm. Overall, the special rater was able to find information on most factors for all of the file formats and effectively apply this information to rating the file formats' endangerment levels.

Extrapolating from all of these findings, I conclude two things. One is that, generally, content encoded in any particular file format is not in as immediate danger of becoming inaccessible as previously believed. That is not to say that information professionals can stop worrying about file formats, nor that it will be easy to access digital content stored in any file format for years to come. What these results indicate is that the situation might not be as dire and irrevocable as the digital preservation community original thought. Nonetheless, professionals still need to take action to maintain access to digital content. These results suggest that it is less about whether or not a file format will become obsolete than it is about how *difficult* it will be to access content encoded in particular file formats over time.

Knowing in advance which file formats are already becoming difficult to access (i.e., if rendering software is not available), and for which rendering software may be difficult to reconstruct (i.e. whether there are specifications available and/or no community or third party support available to reconstruct it), could do much to prevent file formats from becoming more difficult to access over time.

My second conclusion extrapolated from these findings is that a file format endangerment index could be constructed from a small set of factors. Of the lists of 138 factors I found in the literature, I selected three – *rendering software available*, s*pecifications available*, and *community/3rd party support* – to be the most promising candidates for further testing. Contending with the previously large number of ill-defined factors in the proposed automated systems has proven to be entirely infeasible. Reducing the number of factors to only three substantially reduces the complexity of monitoring file format endangerment over the long-term, and most importantly, finally makes it possible.

5.7. Limitations and Challenges

In this section, I address a number of limitations and challenges present in this research. This research was designed to address and clarify several ambiguous aspects of digital preservation research. Due to the ambiguous and often controversial nature of the topics explored in this research, some misunderstandings and confusion occurred in the research process. I describe the questions that arose throughout the process and how I addressed them. I also discuss some of the general limitations of the research design and justify my choices in the face of these limitations.

5.7.1. Questions about Formats and Factors

Throughout the Delphi questionnaire process, during the recruitment period, and after the Delphi studies were complete, I received comments from recruits and participants about the nature of some of the formats in the rating list, the nature of obsolescence, and the definition of file format endangerment. All of these comments provided me with an opportunity to consider the assumptions I bring to this research as well as the assumptions reflected in the digital preservation literature.

There was also some apparent confusion about the difference between looking at factors for evaluating a file format for inclusion in a digital collection and looking at factors as a cause of file format endangerment. Additionally, there was some minor confusion among the factor rating Delphi participants around the term "cause" and the lack of clarity about its relationship with the factors listed.

5.7.2. Programming and Scripting Languages as File Formats

The list of file formats I presented to participants included several different types of programming language source code. I made this decision based primarily on the definition of file formats I used to guide this research: "the internal structure and/or encoding of a file which allows it to be interpreted or rendered in human accessible form" (The National Archives, 2005, p. 8). I included programming and scripting languages because they are encoded in a particular way that is dictated by the rules of the language.

Because programming language source code is written in, and thereby encoded in plain text files, this poses the particularly challenging question of which of the two – the language or the text file format – does one rate when evaluating endangerment levels? When considering it as a text file, it clearly has a very low endangerment level. When considering it as whatever programming or scripting language it is written in, the endangerment level may be different.

Both the format rating Delphi participants and the special rater commented on the different possibilities of rating the programming and scripting languages. In particular participants commented on PHP Source Code files, Perl Script files, C Source Code files, and Fortran Source files. Hypertext Markup Language, Javascript, Cascading Style Sheets, and Extensible Markup Language files are also encoded as plain text and posed similar problems to the Delphi participants. In 44 out of 80 possible instances (55%), participants indicated that they rated them mostly based on the fact that they were written in text files and were human readable once rendered as text. Two participants did consistently explain how and why the format as a programming or scripting language could still be compiled, interpreted, or executed in the time-periods referenced in their ratings. This is an interesting distinction that can benefit from further examination in future research.

5.7.3. Container Formats

Format rating Delphi participants noted some confusion when rating container formats, or single formats that contain or facilitate the streaming of content encoded in a set of possible multimedia codecs. I included several container formats in the format rating Delphi: Real Media, Google Earth Placemark File, Windows Media Video File, Audio Video Interleave File, Apple Quicktime Movie, and MPEG Video File. Participants noted the complexity of rating these container formats in their justification text.

5.7.4. The Nature of Obsolescence and Format Rating Instrument Design

I received comments from recruits and participants before, during, and after the Delphi studies on whether or not file format obsolescence was an actual possibility. Some recruits declined to participate in the study because they did not believe a file format would ever become obsolete. Some of those who participated in the study noted this as well, but reflected this belief in their file format ratings.

Two participants contacted me and expressed a concern over the lack of a file format rating choice for indicating that information encoded in a particular file format will never be inaccessible. I instructed them to select the 20+ years choice and to use the comments section to explain that they did not believe the format would become inaccessible in any timeframe. These findings suggest that it could be beneficial to shift away from addressing the issue in terms of obsolescence and toward a more relative sense of inaccessibility. Taking this into consideration as part of the design of the file format rating instrument--for example, including a rating level indicating that they did not believe that content encoded in a particular format would ever become inaccessible--could have elicited more accurate ratings.

5.7.5. Factors as Causes of Endangerment versus Criteria for Inclusion in a Collection

In the first round of rating, several participant comments indicated that they were rating the factors based on whether or not a collecting institution should collect files encoded in that particular format. During the second round, most of these participants adapted to rating the factors based on their relevancy as causes versus relevancy for inclusion in digital collections. This is most likely the result of both my contacting certain participants to clarify the question, and the result of reading the other participants' responses. Nonetheless, some participants were clearly rating the factors as indicators of suitability for collections even in the second round. Most recognized that the question asked them to rate the factors based on the factors relevancy as a *cause* of file format endangerment.

5.7.6. Cause as Absence or Presence of a Factor

During the factor rating Delphi, some participant answers indicated that there was a lack of clarity around the term “cause” and its relationship with the factors being rated. For example, for the factor *specifications available* some participants noted that the availability of specifications was the opposite of a cause of file format endangerment. To address this problem, I changed the text of the question in the third round to read, “select an answer that indicates how relevant the factor (either its presence or absence) is to indicating a cause of file format endangerment.” It appeared, however, that this lack of clarity did not affect the way participants rated the factors. For example, when participants noted that the factor, *specifications available* was the opposite of a cause of file format endangerment, they still rated it as *very relevant*.

5.7.7. Procedural Questions and the Definition of File Format Endangerment

In the course of responding to the second round questionnaire, one participant proposed his own definition of file format endangerment and an associated causal model with the factors based on his proposed definition. Because the study did not move on to a third round of rating for the initial 21 factors, none of the comments were shared with participants for these factors between the second and third rounds. I attempted to, but found it impossible to isolate information on the seven additional factors from the one, large comment the participant left in the *backward/forward compatibility* comment text box. Consequently, this participant’s comment was not included in the compiled comment document I shared with participants to inform a second round of rating the seven new factors.

However, I believe it is important to discuss a fundamental issue raised by this participant in proposing the alternative definition of file format endangerment. The proposed definition as described by the participant was the following:

> *Format endangerment.* Format endangerment does not have a widely-used common definition, but seems to be used roughly to mean that the format at high risk of becoming obsolete in the near future. In brief abstract terms, a format is "obsolete" when it can no longer be accessed. More specifically, a format should be characterized as "obsolete" when a designated community can no longer extract the significant semantic information which is contained information objects that are encoded in that format. To state this abstract definition in operational terms: a format is obsolete when there are no readily available/operable software that will reliably render files in that format.

By this definition, file format endangerment is the risk a file format exhibits of becoming obsolete, where obsolete is “when there are no readily available/operable software that will reliably render files in that format.” In this case, file format endangerment is defined by the availability of rendering software. Examining the guidelines for creating a formative measure, discussed in detail in section 2.6. above, the formative indicators that are used to measure the designated construct do, in fact, define the construct being measured. The definition proposed by this participant affectively addresses this aspect of formative measure construction and is completed by the inclusion of the factors *available specifications* and *community/3rd party support*, which are primary factors the participant identified in the associated proposed causal model.

Based on the guidelines for creating a formative measure, this participant’s proposed alternative definition, and the results of this study, a final definition for file format endangerment would be: The possibilty that information stored in a particular file format will not be interpretable or renderable in human accessible form due to lack of available rendering software, available specifications, and community/3rd party support.

5.7.8. Implications of Removed File Formats

One limitation of this study is that even with a carefully selected set of experts, there were still some file formats that participants did not believe they had enough experience to rate. While there were no formats that received no ratings, it is somewhat surprising that there were some that more than half of participants did not rate. Further research can benefit from additional exploration into these file formats as well as targeted inquiry in which people with domain expertise, such as video format experts, are asked to make assessments.

5.7.9. Data Sparseness

A general limitation of this study is both a limitation and a rationale for my research. This is the fact that there is no single, definitive source of information on file format endangerment levels to inform expert decisions. While the Delphi method is designed to address issues of sparse data, I believe that stronger pools of existing data on file format risks would make the results of this study stronger.

5.7.10 NARA Corpus as Data Source

An additional limitation of this study is the use of the NARA corpus as a source of file formats. Both the type and count of formats present in the NARA corpus are not necessarily representative of the variety and distribution of file formats in other repositories. Additionally, the NARA corpus does not supply format version information with its file extension information; the format versions rated by study participants are not necessarily the format version(s) present in the NARA collection.

CHAPTER 6: CONCLUSIONS AND FURTHER RESEARCH

Beyond answering the research questions posed here, the findings of this research point to a different way of thinking about file formats in digital preservation. Until recently, the literature has regarded file formats as a substantial threat to the continued accessibility of digital content. Only a few voices, most notably David Rosenthal (2010) and Chris Rusbridge (2006), have said otherwise. When asked to rate how important a risk factor is file format endangerment to the future use of digital materials, participants indicated as a whole that it was important; though notably, they did not indicate that it is *very important*. Even though the file format rating participants rated it as slightly less important than the whole group, they still indicated that they believed it was important. Nonetheless, when asked to rate individual file formats on an endangerment scale, they rated the majority of the file formats as being accessible for 20 years or more. So, even though they believe that file format endangerment is important for digital preservation, they did not determine the selected file formats to be very endangered.

The findings of the file format endangerment-rating exercises taken together with the comments and correspondence from participants and recruits point also to the need to reconsider how file format risk is approached. Thinking in terms of obsolescence is becoming less useful as a way to formulate risk. A number of the study participants and recruits stated that they believed the community has the technological skill and means to maintain access to digital content indefinitely.

The question then becomes not whether a file format will become obsolete, but rather, 1) the ability of the available rendering software's to faithfully represent what was originally intended, and 2) the effort and cost required to maintain and/or create software that can continue to faithfully render the content. This is not a new idea, but one that appears to be getting more traction in the digital preservation community. Rosenthal said in 2010, "Thus the practical questions about the obsolescence of the formats used by today's readers are really how convenient it will be for the eventual reader to access the content, and how much will be spent when in order to reach that level of convenience" (p. 207).

The findings of this study also call into question the notion that assessing file format risk should involve complicated models with dozens of calculated and weighted evaluation factors. As evidenced in the literature, many people have subscribed to this idea that the more factors in an evaluation model, the more accurate the measure will be.

A conversation started by Johan van der Knijff (2013a; 2013b) on the Open Planets Foundation website points out that many of the factors included in these models are theoretical, untested, and sometimes not testable.

The findings of this research suggest that most of the file format evaluation factors are not primary or even secondary causes of file format endangerment. After the factor rating Delphi was complete, one participant wrote to me noting how surprised she was at how few factors she thought were actual causes of endangerment.

6.1. Further Research

The findings reported here have implications for several avenues of future research. The most obvious avenues are 1) continued study of file format endangerment levels centered on the file formats removed from the study and other lesser known file formats, 2) continued development and refinement of the proposed file format endangerment index, 3) operationalization of the file format endangerment index within an early warning system, and 4) exploration into how a file format endangerment early warning system can trigger decisions and actions within the digital preservation and memory institution environment.

6.1.1. Continued File Format Evaluation

The seven file formats that were removed from the study provide an opportunity for further exploration into their nature and yet established file format endangerment levels. A duplicate Delphi study could be performed with these formats but with participants who know enough about the file formats to rate them. Additionally, information could be located and scored for these formats of the file format endangerment index factors. This process could be repeated for other lesser-known formats, specifically those that are used within smaller domains such as the hard sciences.

Once methods are established for applying the file format endangerment index to assessing file format endangerment, these methods could be applied to all known file formats. Ideally, systematic data collection for the index and sharing would take place within a loosely organized federation of institutions and individuals, similar to how data collection and evaluation is performed for the International Union for the Conservation of Nature (IUCN) Red List of Threatened Species (2013).

6.1.2. Further Index Development

The research discussed here is the first step toward creating a file format endangerment index that can be used to detect when content encoded in a particular file format may be more difficult to access. Following the recommendations of Diamantopoulos and Winklehofer (2001) for constructing an index, the next steps are to test and validate the index. Testing and validating an index first requires that data be collected for the selected formative indicators.

A starting point for data collection can be to use the data collected by the special rater (See *Appendix E: File Format Factor Information Guide*) and the data collection suggestions provided by the factor rating Delphi participants. From there, appropriate tests for collinearity could be performed, and the index can be validated against the file format ratings collected in the format rating Delphi study and from future collected expert ratings. From there, continued data collection for each of the factors could be conducted in conjunction with continued assessment of the collected data.

Once the factors selected for the index have been adjusted and validated, the measure could be put to immediate use in evaluating file format endangerment levels both in the local and global contexts. Coordination of cooperative efforts with institutions, coalitions, and other researchers who are working in this area could expand data collection and the application of the index.

Once the primary factors have been tested established, it would be valuable to explore nuances of each of the factors. For example, the factor, *specifications available*, could be examined not just by whether or not specifications are available, but by how useful the specifications are to the creation or recreation of viable rendering software. Additionally, the factor, *rendering software available*, could be evaluated not just for whether or not it is available, but how faithfully it represents the original intended representation of the encoded content.

6.1.3. File Format Endangerment Early Warning System

Further research may begin around the development of early warning systems for file format endangerment, following on the models developed for use in early warning systems used in detecting epidemics, natural disasters, societal conflicts, terrorist attacks, and other threats to humanity. Basher defined 'early warning' as "the provision of information on an emerging dangerous circumstance where that information can enable action in advance to reduce the risks involved" (2006, p. 2168). The United Nations International Strategy for Disaster Reduction (UN ISDR) defined early warning as, "the provision of timely and effective information, through identified institutions, that allows individuals to take action to avoid or reduce their risk and prepare for effective response" (UN ISDR, 2006, p. 2). Choo defined an early warning system as a "network of actors, practices, resources, and technologies that has the common goal of detecting and warning about an imminent threat so that preventive measures can be taken to control the threat or mitigate its harmful effect" (2009, p. 1072).

By using these definitions, drawing parallels with file format endangerment assessment is straightforward: all involve collecting and providing information on potential risk that could inform decisions and actions that may reduce or avoid the risk. There is an opportunity in applying the methods defined in the early warning system development area to the file format endangerment research area.

6.1.4. Triggered Actions and Decision-Making

Once a warning is signaled from an early warning system, it would be valuable to have a framework in place to guide appropriate actions and decision-making that will be triggered from the warning; particularly for use within memory institutions. A number of scenarios could be investigated to establish what kinds of decisions the appropriate stakeholders would face. These scenarios could include, but not be not limited to making cost-benefit analysis decisions on whether or not action can or should be taken to maintain access to content encoded in vulnerable file formats. The scenarios could also include paths for deciding which strategy can be used to access content encoded in the vulnerable formats or may connect the stakeholder with communities that can offer up to date guidance to make these decisions. This application further makes the case for continued efforts to collect information about file formats, particularly migration paths and emulation strategies, as well as use-cases for accessing content encoded in particular file formats, and more structured and accessible communities to which decision-makers may turn for advice when necessary.

APPENDIX A: RECRUITMENT TEXT

Recruitment Text for Delphi Participants

Dear ______________________,

I am writing to invite you to participate in my study on file format endangerment. In this study, I am harnessing pooled expertise to 1) establish a baseline endangerment level for 100 file formats, and 2) choose factors for a file format endangerment index. I have identified you as an expert in this area, i.e. someone who has demonstrated knowledge about and/or experience with file formats in a digital preservation context and the factors that may or may not cause them to be endangered.

There has been much discussion over the years about whether and how file formats affect the preservation of digital information over time. Currently, there are no tested, scientific methods to collect and analyze data that can answer these questions. To address this, I will develop methods to systematically determine to what degree the format in which a digital file is saved poses a threat to continued access of the information it contains. The first steps of this process are to establish a baseline understanding of current levels of file format-related risk, and to create a valid measure, or index, to guide future data collection and analysis.

What do I mean by file format endangerment? For the purposes of this study, a file format is, "the internal structure and/or encoding of a file which allows it to be interpreted or rendered in human accessible form" (The National Archives, 2005, p. 8). File format endangerment indicates the possibility that information stored in a particular file format will not be interpretable or renderable in human accessible form within a certain timeframe. You may be more familiar with the term, "obsolete" which indicates the state in which a file format is no longer in common use or is no longer easily accessible. I propose the use of the term "endangerment" to refer to the stages leading up to obsolescence, much as living species that are at risk of extinction are labeled as endangered.

The National Archives. (2005, July). The Selection of Preservation Formats. Retrieved May 8, 2013 from http://longtermdata.com/pdfs/Fresko_TNASelection.pdf.

What will I ask you to do? I will ask you to participate in one of two separate Delphi questionnaires where I will ask you a series of questions over several rounds. If you agree to participate, I will first ask you to complete a short questionnaire that asks you about experience you have had in managing and evaluating file formats in a digital preservation environment and/or conducting research on file formats in a digital preservation context. Based on this information, I will select the Delphi group in which I will ask you to participate.

In one questionnaire group, I will ask you to rate 100 file formats on a 6-point endangerment scale and to briefly explain your rating for each format. In the second questionnaire group, I will also ask you to review and vote on a list of factors for possible inclusion in a file format endangerment index. After completing one of these questionnaires, I will ask you to review the anonymous answers and rationale of your fellow study participants, to reconsider your answers in light of fellow participants' answers, and to re-answer the questions in additional questionnaire rounds.

Throughout this process, you will remain anonymous to the other participants. Your and the other participants' names may be shared in publications about this study to lend credibility to the results and to acknowledge you for your contribution. Specific comments collected during the questionnaire process may be used in publications, but they will not be associated with you as an individual participant.

I anticipate that this study will take 5-10 hours of your time over the course of 4-8 weeks starting September 30th, 2013.

What are the benefits of this research? This study will benefit the digital preservation discipline by contributing to a greater understanding of file format endangerment levels. Most importantly, this research will contribute to the creation of a file format endangerment index that can be applied the scientific assessment of real and perceived risks of file formats in digital preservation. By participating in this study, you will have the opportunity to share your insights as well as learn from others' experiences with file formats in digital preservation. This experience may benefit you personally as well as the digital preservation community by expanding the conversation about file format endangerment.

I value your expertise in this area and I believe that your participation in this study will increase its strength and integrity. Are you interested in participating? If so, please respond to this email indicating that you would like to participate. I will provide you with detailed instructions on how to access the first, short expertise questionnaire within the next few days.

Additionally, if you know of other experts in this area who you believe I should invite, please let me know who they are and how I can contact them.

Kind regards,

Heather Ryan
PhD Candidate
School of Information & Library Science
University of North Carolina at Chapel Hill
heather@longtermdata.com OR hbowden@email.unc.edu
Advisor: Dr. Cal Lee

Recruitment Text B, for Special Reviewer

Dear ______________________,

I am writing to invite you to participate in my research on file format endangerment. Through my research, I am harnessing pooled expertise to 1) establish a baseline endangerment level for 50 test file formats, and 2) choose factors for a file format endangerment index. I have identified you as an expert in this area, i.e. someone who has demonstrated knowledge about and/or experience with file formats in a digital preservation context and the factors that may or may not cause them to be endangered.

There has been much discussion over the years about the questions of whether and how file formats affect the preservation of digital information over time. Currently, there are no tested, scientific methods to collect and analyze data that can answer these questions. To address this, I will develop methods to systematically determine to what degree the format in which a digital file is saved poses a threat to continued access of the information it contains. The first steps of this process are to establish a baseline understanding of current levels of file format-related risk, and to create a valid measure, or index to guide future data collection and analysis.

What do I mean by file format endangerment? For the purposes of this study, a file format is, "the internal structure and/or encoding of a file which allows it to be interpreted or rendered in human accessible form" (The National Archives, 2005, p. 8). File format endangerment indicates the probability that information stored in a particular file format will not be interpretable or renderable in human accessible form in 20 years. Most commonly, the term, "obsolete" indicates the state in which a file format is no longer in common use or is no longer easily accessible. I propose the use of the term "endangerment" to refer to the stages leading up to obsolescence, much as living species that are at risk of extinction are labeled as endangered.

The National Archives. (2005, July). The Selection of Preservation Formats. Retrieved May 8, 2013 from http://longtermdata.com/pdfs/Fresko_TNASelection.pdf.

What will I ask you to do? I will ask you to participate in a questionnaire where I will ask you to collect and share specific information about a list of up to 50 file formats and rate their level of endangerment based on the information you collected. Before the questionnaire, I will ask you to participate in a training session where I will review the challenges associated with file formats and their role in digital preservation, I will review the type of information I would like you to collect for each file format, and I will provide you with a guide on where you might collect the file format information. After the questionnaire, I will ask you to participate in a one-on-one interview where I will ask you about the process you used to collect the file format information, how you applied the information to rating the file formats, and how well each specific type (factor) of information helped you rate the file formats.

Your identity will not be revealed in any future publications about this research. Specific comments collected during the questionnaire process may be used in publications, but they will not be associated with your name.

I anticipate that this study will take approximately 27 hours of your time over the course of 3 weeks. The initial training should take one hour, the questionnaire will take approximately 25 hours, and the post-questionnaire interview will take approximately 1 hour.

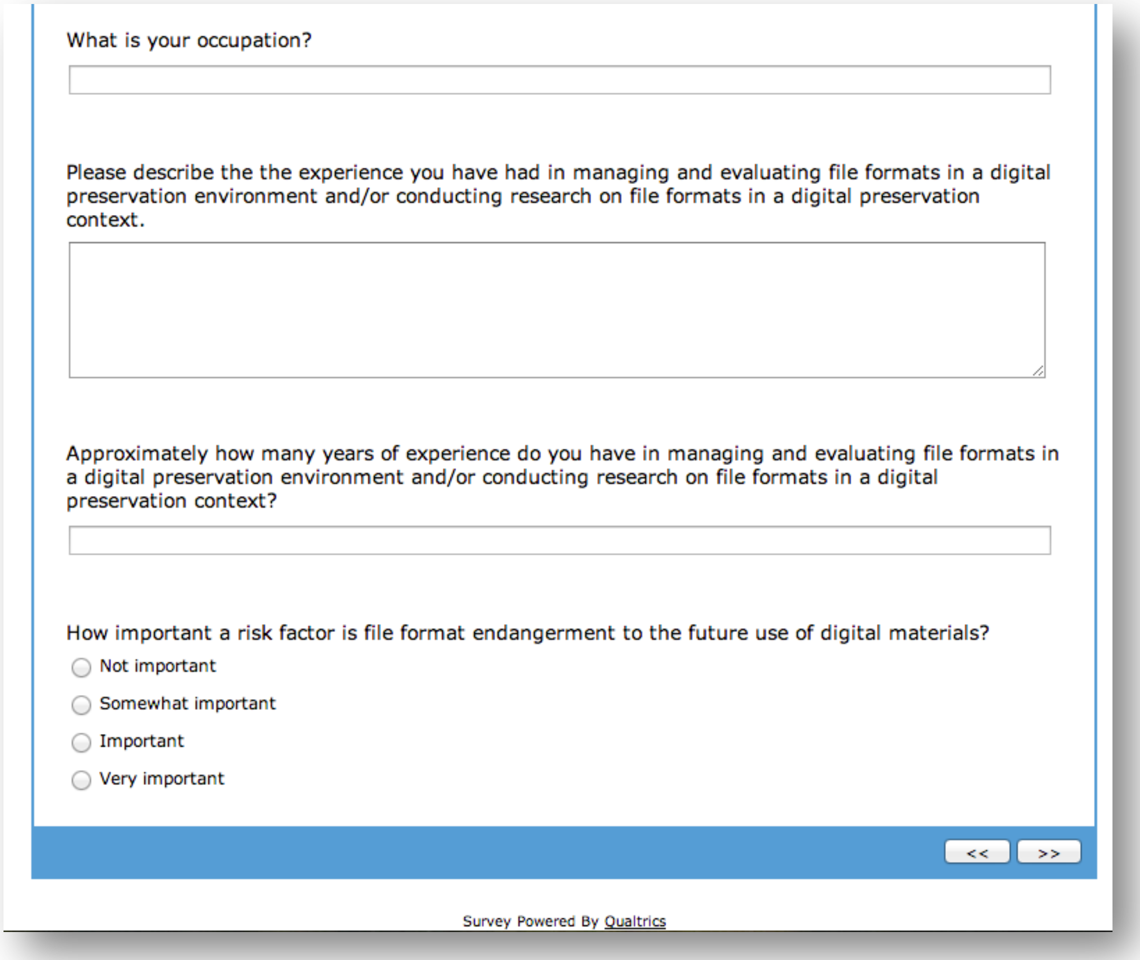
What are the benefits of this research? This study will benefit the digital preservation discipline by contributing to a greater understanding of file format endangerment levels, and by providing valuable information that will contribute to the creation of a file format endangerment index.

Are you interested in participating? If so, please respond to this email indicating that you would like to participate. I will provide you with detailed information on how we will proceed.

APPENDIX B: QUESTIONNAIRE DESIGNS

The following images are sample sections of the four questionnaires I used in this research. In the first questionnaire, I asked questions about participants' number of years and type of experience working with file formats in a digital preservation context. In the second questionnaire, I asked the participant to rate endangerment levels for the test file formats and explain their ratings. In the third questionnaire I asked the participant to choose the degree to which he or she believed each of presented factors are relevant formative indicators for a file format endangerment index. There was also space for participants to recommend factors that were not included in the original list. In the fourth questionnaire, I asked the participant to collect data for each file format evaluation factor listed, for each file format. I asked the participant to rate each listed file format based on the data collected, and I also asked the participant to rate the factors for relevancy as a cause of file format endangerment. The sample pages presented here show only one file format of the entire list, and only one factor of the list that will be presented to participants.

Questionnaire 1: Experience with File Formats



What is your occupation?

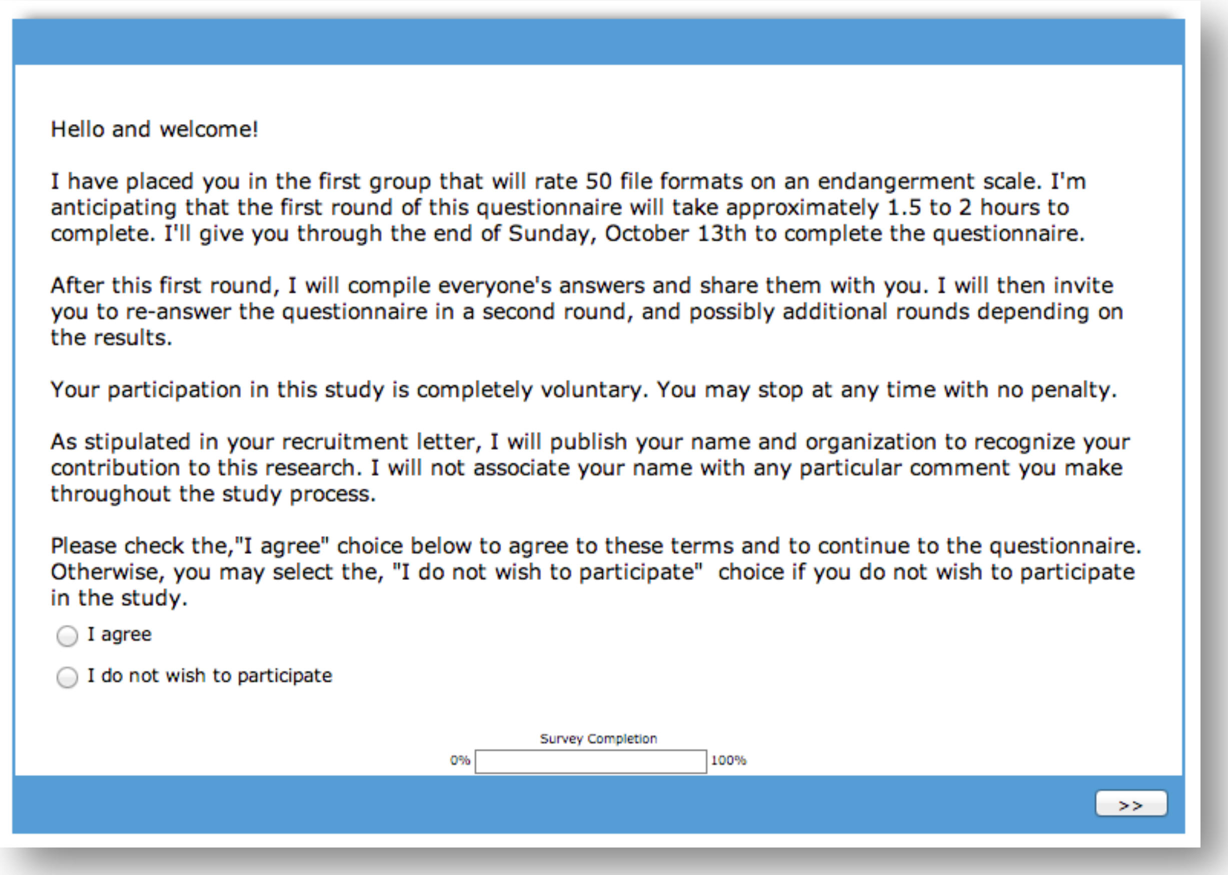
Please describe the the experience you have had in managing and evaluating file formats in a digital preservation environment and/or conducting research on file formats in a digital preservation context.

Approximately how many years of experience do you have in managing and evaluating file formats in a digital preservation environment and/or conducting research on file formats in a digital preservation context?

How important a risk factor is file format endangerment to the future use of digital materials?

- ○ Not important
- ○ Somewhat important
- ○ Important
- ○ Very important

Questionnaire 2: Rating File Formats



Hello and welcome!

I have placed you in the first group that will rate 50 file formats on an endangerment scale. I'm anticipating that the first round of this questionnaire will take approximately 1.5 to 2 hours to complete. I'll give you through the end of Sunday, October 13th to complete the questionnaire.

After this first round, I will compile everyone's answers and share them with you. I will then invite you to re-answer the questionnaire in a second round, and possibly additional rounds depending on the results.

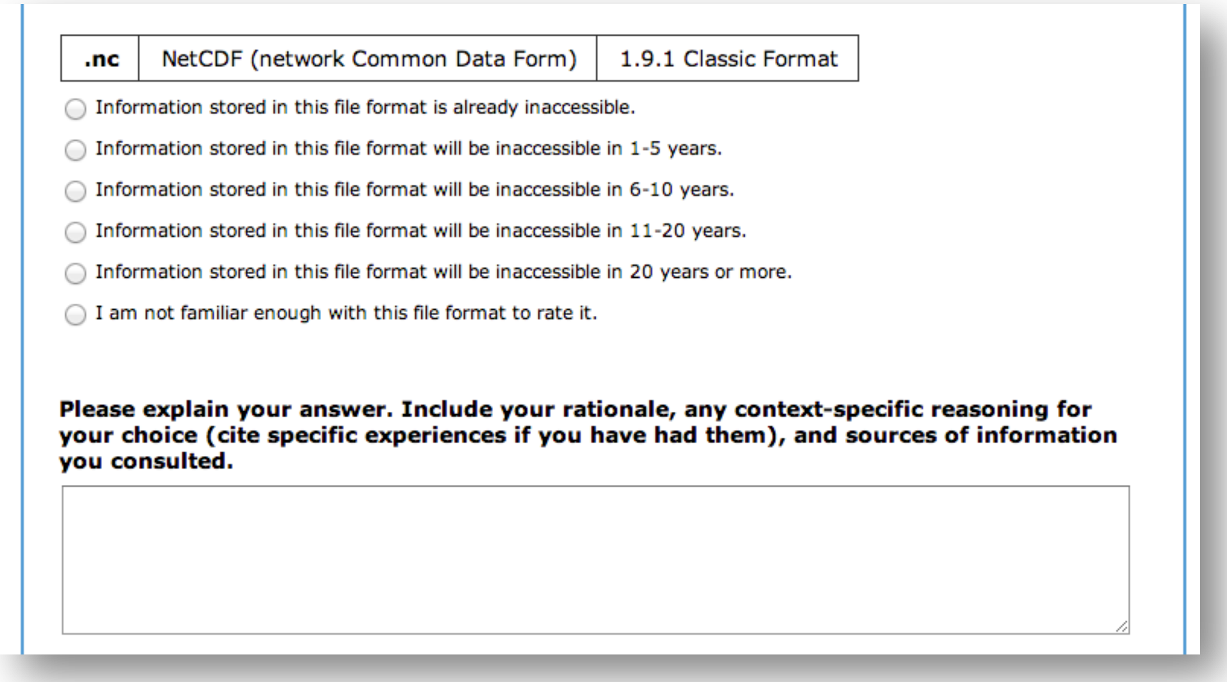
Your participation in this study is completely voluntary. You may stop at any time with no penalty.

As stipulated in your recruitment letter, I will publish your name and organization to recognize your contribution to this research. I will not associate your name with any particular comment you make throughout the study process.

Please check the,"I agree" choice below to agree to these terms and to continue to the questionnaire. Otherwise, you may select the, "I do not wish to participate" choice if you do not wish to participate in the study.

○ I agree

○ I do not wish to participate



.nc	NetCDF (network Common Data Form)	1.9.1 Classic Format

○ Information stored in this file format is already inaccessible.

○ Information stored in this file format will be inaccessible in 1-5 years.

○ Information stored in this file format will be inaccessible in 6-10 years.

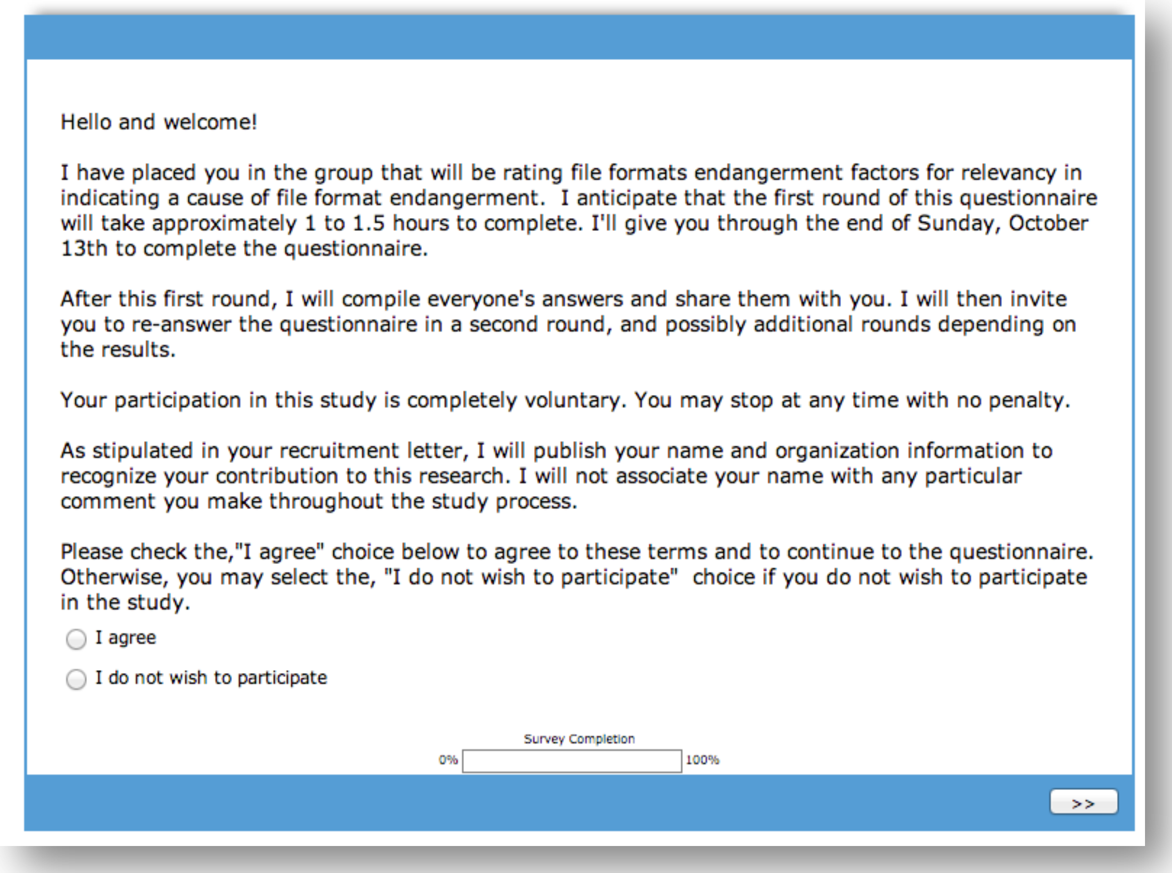
○ Information stored in this file format will be inaccessible in 11-20 years.

○ Information stored in this file format will be inaccessible in 20 years or more.

○ I am not familiar enough with this file format to rate it.

Please explain your answer. Include your rationale, any context-specific reasoning for your choice (cite specific experiences if you have had them), and sources of information you consulted.

Questionnaire 3: Factor Rating



Hello and welcome!

I have placed you in the group that will be rating file formats endangerment factors for relevancy in indicating a cause of file format endangerment. I anticipate that the first round of this questionnaire will take approximately 1 to 1.5 hours to complete. I'll give you through the end of Sunday, October 13th to complete the questionnaire.

After this first round, I will compile everyone's answers and share them with you. I will then invite you to re-answer the questionnaire in a second round, and possibly additional rounds depending on the results.

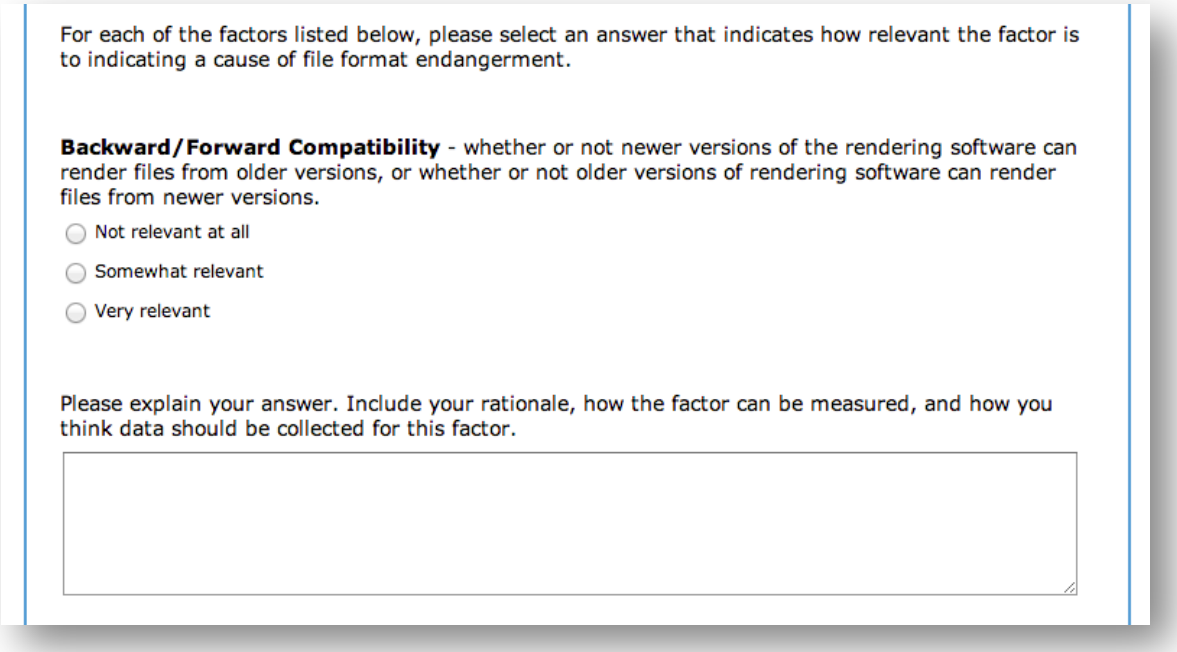
Your participation in this study is completely voluntary. You may stop at any time with no penalty.

As stipulated in your recruitment letter, I will publish your name and organization information to recognize your contribution to this research. I will not associate your name with any particular comment you make throughout the study process.

Please check the,"I agree" choice below to agree to these terms and to continue to the questionnaire. Otherwise, you may select the, "I do not wish to participate" choice if you do not wish to participate in the study.

○ I agree

○ I do not wish to participate



For each of the factors listed below, please select an answer that indicates how relevant the factor is to indicating a cause of file format endangerment.

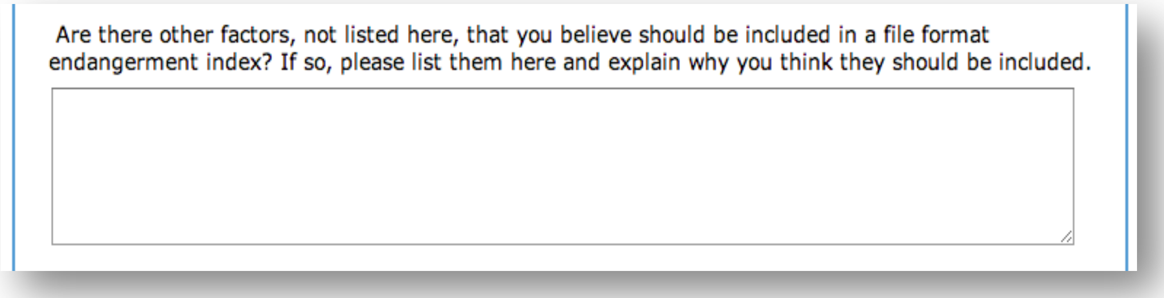
Backward/Forward Compatibility - whether or not newer versions of the rendering software can render files from older versions, or whether or not older versions of rendering software can render files from newer versions.

○ Not relevant at all

○ Somewhat relevant

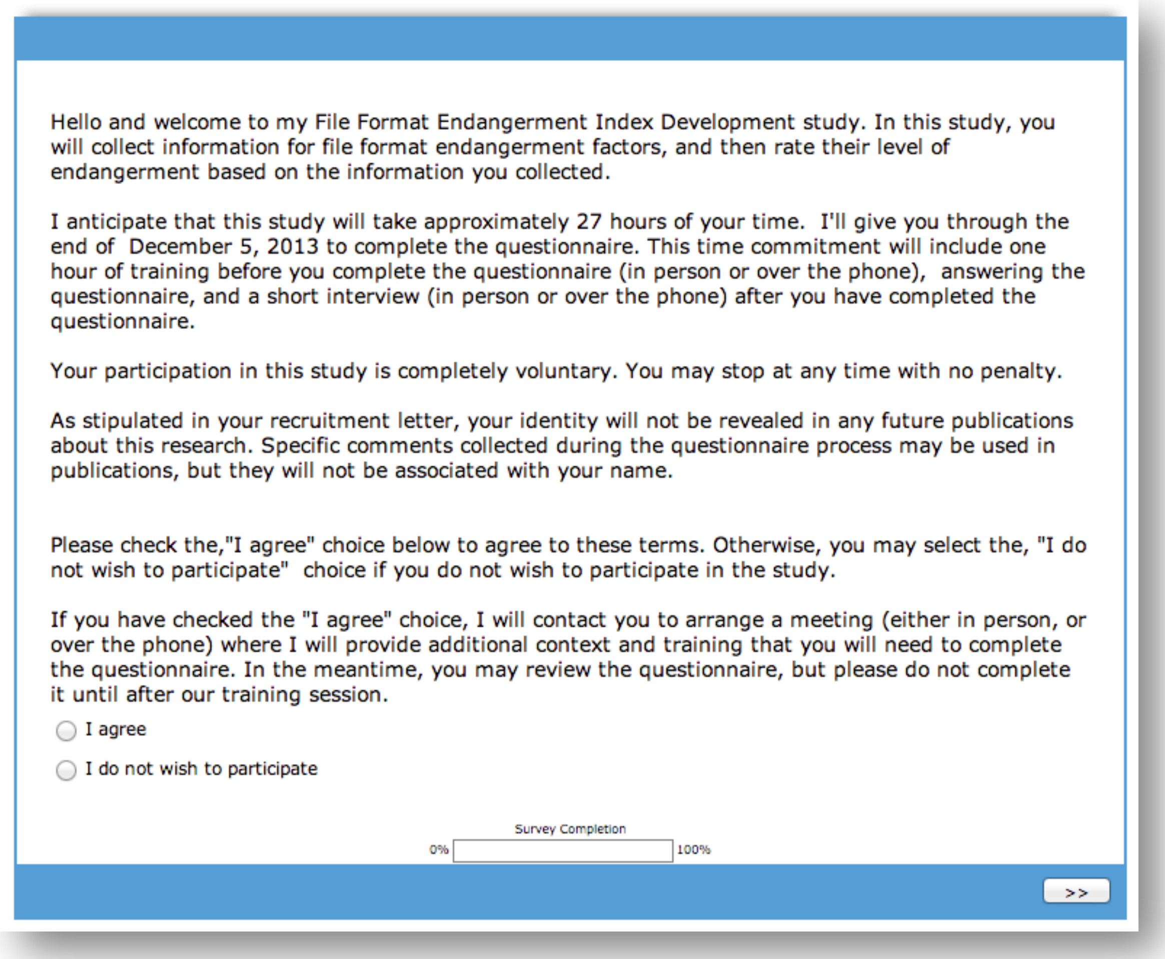
○ Very relevant

Please explain your answer. Include your rationale, how the factor can be measured, and how you think data should be collected for this factor.



Are there other factors, not listed here, that you believe should be included in a file format endangerment index? If so, please list them here and explain why you think they should be included.

Questionnaire 3: Factor Testing



Hello and welcome to my File Format Endangerment Index Development study. In this study, you will collect information for file format endangerment factors, and then rate their level of endangerment based on the information you collected.

I anticipate that this study will take approximately 27 hours of your time. I'll give you through the end of December 5, 2013 to complete the questionnaire. This time commitment will include one hour of training before you complete the questionnaire (in person or over the phone), answering the questionnaire, and a short interview (in person or over the phone) after you have completed the questionnaire.

Your participation in this study is completely voluntary. You may stop at any time with no penalty.

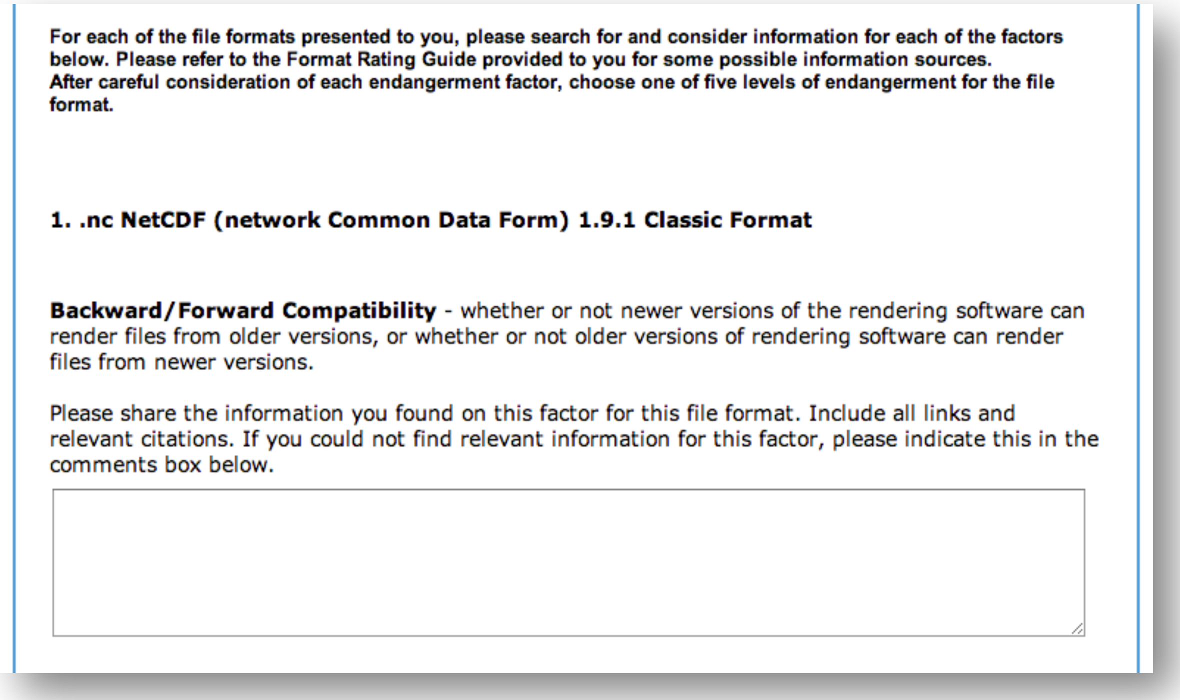
As stipulated in your recruitment letter, your identity will not be revealed in any future publications about this research. Specific comments collected during the questionnaire process may be used in publications, but they will not be associated with your name.

Please check the,"I agree" choice below to agree to these terms. Otherwise, you may select the, "I do not wish to participate" choice if you do not wish to participate in the study.

If you have checked the "I agree" choice, I will contact you to arrange a meeting (either in person, or over the phone) where I will provide additional context and training that you will need to complete the questionnaire. In the meantime, you may review the questionnaire, but please do not complete it until after our training session.

○ I agree

○ I do not wish to participate

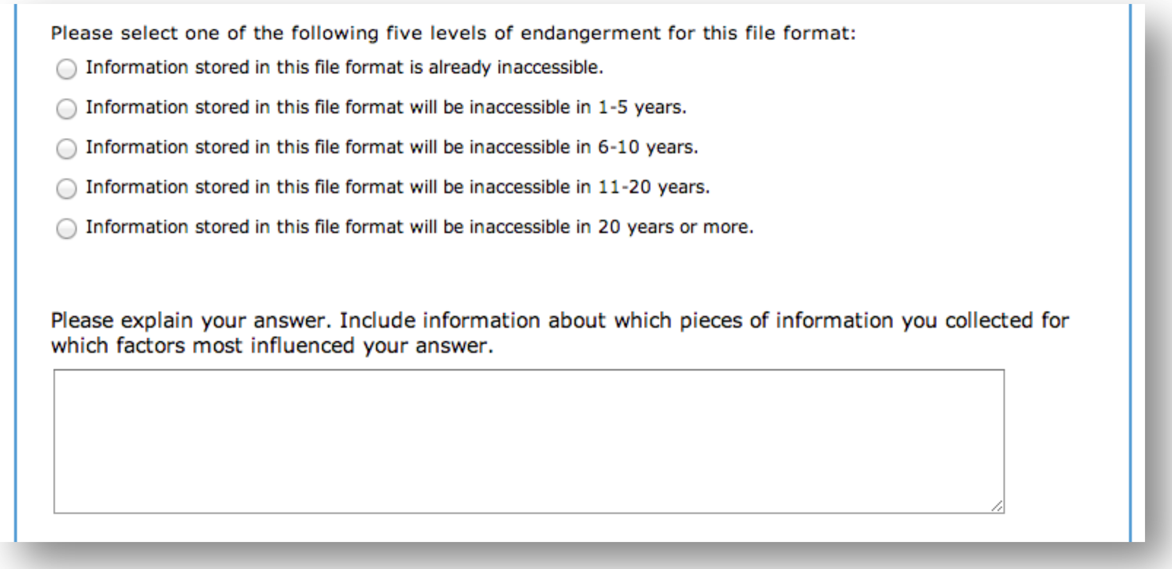


For each of the file formats presented to you, please search for and consider information for each of the factors below. Please refer to the Format Rating Guide provided to you for some possible information sources.
After careful consideration of each endangerment factor, choose one of five levels of endangerment for the file format.

1. .nc NetCDF (network Common Data Form) 1.9.1 Classic Format

Backward/Forward Compatibility - whether or not newer versions of the rendering software can render files from older versions, or whether or not older versions of rendering software can render files from newer versions.

Please share the information you found on this factor for this file format. Include all links and relevant citations. If you could not find relevant information for this factor, please indicate this in the comments box below.



Please select one of the following five levels of endangerment for this file format:

○ Information stored in this file format is already inaccessible.

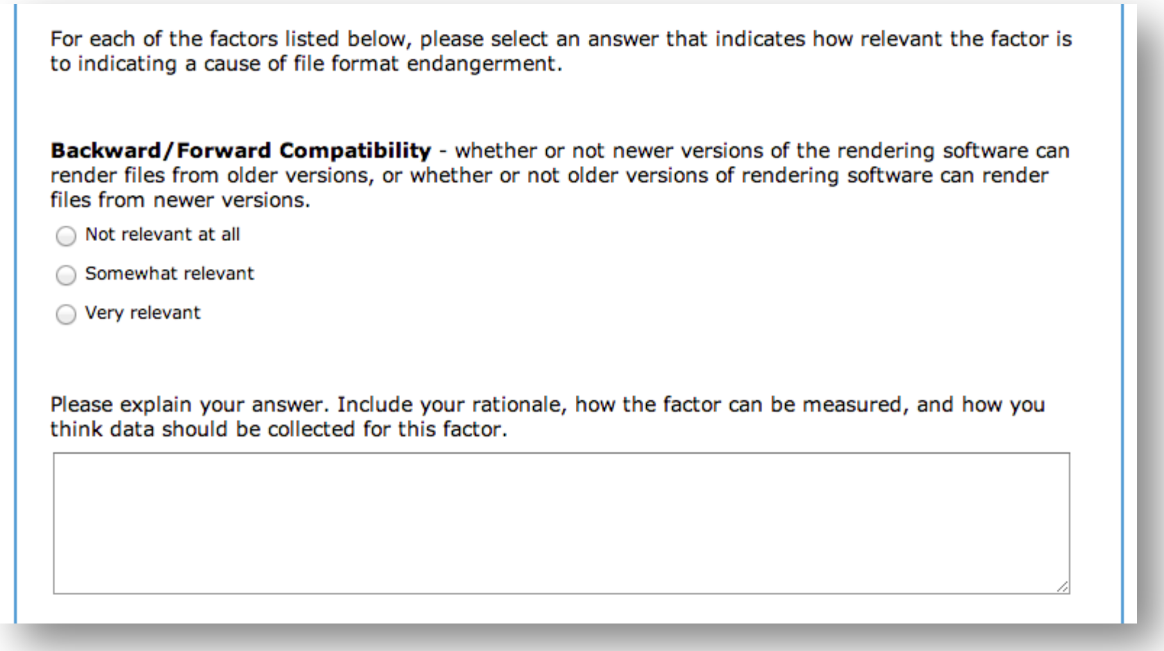
○ Information stored in this file format will be inaccessible in 1-5 years.

○ Information stored in this file format will be inaccessible in 6-10 years.

○ Information stored in this file format will be inaccessible in 11-20 years.

○ Information stored in this file format will be inaccessible in 20 years or more.

Please explain your answer. Include information about which pieces of information you collected for which factors most influenced your answer.



For each of the factors listed below, please select an answer that indicates how relevant the factor is to indicating a cause of file format endangerment.

Backward/Forward Compatibility - whether or not newer versions of the rendering software can render files from older versions, or whether or not older versions of rendering software can render files from newer versions.

○ Not relevant at all

○ Somewhat relevant

○ Very relevant

Please explain your answer. Include your rationale, how the factor can be measured, and how you think data should be collected for this factor.

APPENDIX C: FILE FORMAT RATING GUIDE FOR PARTICIPANTS

In asking you to rate file formats in terms of levels of endangerment, my aim is to create a general assessment of the fifteen file formats' endangerment level. You may have a general feeling of how endangered a file format may be, or you may have had specific experiences or data that inform your ratings. If you have had specific experiences or data that lead you to your conclusions, please cite them in the explanation you provide. If you do not have specific experiences or data, but you have a general understanding of how endangered (or not) a file format is, please state this in your explanation.

Definitions

For the purposes of this study, a **file format** is, "the internal structure and/or encoding of a file which allows it to be interpreted or rendered in human accessible form" (The National Archives, 2005, p. 8). File format version information is not included in this study. Future research will encompass higher degrees of granularity that includes versions, but for now, this study focuses only on general file formats.

File format endangerment indicates the probability that information stored in a particular file format will not be interpretable or renderable in human accessible form 20 years from now.

Inaccessible in this context means that information is not capable of being used or seen.

Reference: The National Archives. (2005, July). The Selection of Preservation Formats. Retrieved May 8, 2013 from http://longtermdata.com/pdfs/Fresko_TNASelection.pdf.

REFERENCES

Abrams, S. & Flecker, D. (2005). *Proposal for a global digital format registry*. Retrieved October, 19, 2008 from http://hul.harvard.edu/gdfr/documents/Proposal-2005-09-29.doc

Abrams, S., & Seaman, D. (2003). Towards a global digital format registry. *World Library and Information Congress: 69th IFLA General Conference and Council,* August 1-9. Berlin: IFLA.

Allison, A., Currall, J., Moss, M., & Stuart, S. (2005). Digital identity matters. *Journal of the American Society of Information Science and Technology, 56*(4), 364-372.

Anderson, R., Frost, H., Hoebelheinrich, N., & Johnson, K. (2005). The AIHT at Stanford University. *D-Lib Magazine*, *11*(12).

Angermeier, P. L. (1995, February). Ecological attributes of extinction-prone species: Loss of freshwater fishes of Virginia. *Conservation Biology, 9*, 143-158.

Archive Team. (2012). *Statement of project*. Retrieved January 24, 2014 from http://fileformats.archiveteam.org/wiki/Statement_of_Project

Arms, C.R., & Fleischhauer, C. (2005). Digital formats: Factors for sustainability, functionality, and quality. *Imaging Science & Technology Archiving 2005*, Washington, DC, April 2005, 222-227.

Australian Partnership for Sustainable Repositories. (2006, September). *AONS system documentation* (Revision 169 2006-09-29). Canberra, Australia: Curtis, J.

Australian Partnership for Sustainable Repositories. (2007, November). *Report of the format notification and obsolescence service (AONS II)*. Canberra, Australia: Pearson, D. & Walker, M.

Babbie, E. (1990). *Survey research methods*, 2nd Ed. Belmont, CA: Wadsworth Publishing Company.

Barve, S. (2007). File formats in digital preservation. In Madalli, D.P, & Madalli, P. (Eds.), *International Conference on Semantic Web & Digital Libraries: ICSD-2007*, 239-248.

Basher, R. (2006). Global early warning systems for natural hazards: Systematic and people-centered. *Philosophical Transactions of the Royal Society A-Mathematical Physical and Engineering Sciences, 364*, 2167-2180.

Bearman, D. (1994a). *Electronic evidence*. Washington, D.C.: Archives & Museum Informatics.

Bearman, D. (1994b). *Virtual electronic junkyard or cultural treasure trove?* Washington, D.C.: Archives & Museum Informatics.

Becker, C., Kulovitz, H. Brown, A. (2008, June). *Planets: Report on service integration in Plato 2*. Project: IST-2006-033789 Planets. Deliverable: PP4/D3.

Bennett, J. C. (1997). *A Framework of Data Types and Formats, and Issues Affecting the Long Term Preservation of Digital Material*. (British Library Research and Innovation Report 50). University of Bath, British Library Research and Innovation Centre.

Boje, D. M., & Murnighan, J. K. (1982, October). Group confidence pressures in iterative decisions. *Management Science, 28*(10), 1187-1196.

Bollacker, K. D. (2010, April). Avoiding a digital dark age. *American Scientist, 98*(2), 106-110.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley-Interscience.

Borghoff, U. M., Rödig, P., Schmitz, L., & Scheffczyk, J. (2006). *Long-Term preservation of digital documents*. Berlin: Springer.

Brown, A. (2005b, April). Automating preservation: New developments in the PRONOM service. *RLG DigiNews, 9.*

Brown, A. (2007, June). Developing practical approaches to active preservation. *International Journal of Digital Curation, 2,* 3-11.

Campbell-Kelly, M., & Aspray, W. (2004). *Computer: A history of the information machine* (2nd ed.). Boulder, CP: Westview Press.

Choo, C.W. (2009). Information use and early warning effectiveness: Perspective and prospects. *Journal of the American Society for Information Science and Technology, 60*, 1071-1082.

Collier, N., et al. (2011). Navigating the information storm: Web-based global health surveillance in BioCaster. In T. Kass-Hout & K. Zhang (Eds.), *Biosurveillance: Methods and case studies* (pp. 291-310). Boca Raton: CRC Press.

Conway, P. (1996). *Preservation in the Digital World*. Washington, D.C.: Council on Library and Information Resources Commission on Preservation and Access.

Consultative Committee for Space Data Systems (CCSDS). (2002, January). *Reference Model for an Open Archival Information System: Blue Book.* Retrieved December 27, 2013 from http://public.ccsds.org/publications/archive/910x4b2e1.pdf

Consultative Committee for Space Data Systems (CCSDS). (2012, June). *Reference Model for an Open Archival Information System: Magenta Book, Issue 2.* Retrieved June 5, 2013 from http://public.ccsds.org/publications/archive/650x0m2.pdf

Dajani, J.S., Sincoff, M.Z., & Talley, W.K. (1979). Stability and agreement criteria for the termination of Delphi studies. *Technological Forecasting and Social Change, 13*, 83-90.

Dalkey, N.C. (1968). Predicting the future. *National Conference on Fluid Power*, Chicago, Illinois.

Delbecq, A.L., Van de Ven, A.H., & Gustafson, D.H. (1975). *Group techniques for program planning*. Glenview, IL: Scott, Foresmann, and Co.

Diamantopoulos, A., Riefler, P., & Roth, K.P. (2008). Advancing formative measurement models. *Journal of Business Research, 61*, 1203-1218.

Diamantopoulos, A. & Winklhofer, H.M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, 38(2).

Digital Preservation Europe. (2007). *DPE Research Roadmap, DPE-D7.2*. Retrieved February 17, 2013 from http://www.digitalpreservationeurope.eu/publications/dpe_research_roadmap_D72.pdf

Doak, D. F., Finkelstein, M. E., & Bakker, V. J. (2009). Population viability analysis. In S. A. Levine (Ed.), *The Princeton Guide to Ecology* (pp. 521–528). Princeton, N.J: Princeton University Press.

Dollar, C.M. (1971). Documentation of machine-readable records and research: A historian's view. *Prologue 3*(1), 27-31.

Ex Libris Group. (2010). *Ex Libris Rosetta: A digital preservation system product description.* Retrieved February 21, 2013 from http://www.exlibrisgroup.com/category/RosettaOverview

Faria, L. (2013). Scout - A preservation watch system. Retrieved December 29, 2013 from the Open Planets Foundation website http://www.openplanetsfoundation.org/blogs/2013-12-16-scout-preservation-watch-system

Faria, L., Akbik, A., Sierman, B., Ras, M., Ferreira, M., & Ramalho, J.C. (2013). Automatic preservation watch using extraction on the web. *Proceedings of the10th International Conference on the Preservation of Digital Objects, Lisbon, Portugal*.

Ferreira, M., Baptista, A.A., & Ramalho, J. C. (2006, July). A foundation for automatic digital preservation. *Ariadne, 48*.

Ferreira, M., Baptista, A. A., & Ramalho, J. C. (2007). An intelligent decision support system for digital preservation. *International Journal on Digital Libraries*, *6*(4), 295-304.

Fischer, T. (2003). LaTeX as an archiving format: Benefits and problems. *Proceedings of the Sixth International Symposium on Electronic Theses and Dissertations ETD2003*. Berlin: Humboldt-Universität zu.

Gerber, L. & González-Suárez, M. (2010) Population viability analysis: Origins and contributions. *Nature Education Knowledge* 3(10):15

Gordon, T. J., & Helmer, O. (1964). *Report on a long-range forecasting study*. Santa Monica, CA: RAND Corporation.

Graf, R. & Gordea, S. (2013). A risk analysis of file formats for preservation planning. *Proceedings of the10th International Conference on the Preservation of Digital Objects, Lisbon, Portugal*.

Harvard University Libraries [HUL]. (2006, October). *Global Digital Format Registry (GDFR) data model* (Version 5.0.14). Harvard: Abrams, S., & Goethals, A.

Harvard University Libraries [HUL]. (2008, May). *Global Digital Format Registry (GDFR) data model specifications* (Draft 5.0.5). Harvard.

Hedstrom, M. (1984). *Archives and manuscripts: Machine-readable records*, Basic Manual Series. Chicago, IL: Society of American Archivists.

Hedstrom, M. (1998). Digital preservation: A time bomb for digital libraries. *Computers and the Humanities, 31*, 189-202.

Hedstrom, M., Lee, C.A., Olson, J.S. & Lampe, C.A. (2006). "The old version flickers more": Digital preservation from the user's perspective. *American Archivist, 69*, 159-187.

Helmer, O., & Rescher, N. (1959). On the epistemology of the inexact sciences. *Management Science, 6*(1), 25-52.

Higgins, S. (2008). The DCC curation lifecycle model. *The International Journal of Digital Curation, 1*(3).

Higgins, S. (2011). Digital curation: The emergence of a new discipline. *The International Journal of Digital Curation, 6*(2), 78-88.

Huc, C., et al. (2004). *Criteria for evaluating data formats in terms of their suitability for ensuring information long term preservation*. Retrieved February 22, 2013 from http://www.docstoc.com/docs/57566655/Criteria-for-evaluating-data-formats.

Hunter, J. & Choudhury, S. (2004). A semi-automated digital preservation system based on semantic web services. In *Global Reach and Diverse Impact: Fourth ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '04), June 7-11, 2011,Tucson, AZ (pp. 268-278). Association for Computing Machinery.

Hunter, J. & Choudhury, S. (2006). PANIC: an integrated approach to the preservation of composite digital objects using semantic web services. *International Journal of Digital Libraries, 6*(2), 174-183.

International Society for Infectious Diseases [ISID]. (2009). *ProMED Mail*. Retrieved May 5, 2009, from http://www.promedmail.org.

International Union for the Conservation of Nature [IUCN]. (2013). Red list overview. Retrieved from http://www.iucnredlist.org/about/red-list-overview#assessment_process.

InterPARES. (2007, March). General study 11 final report: Selecting digital file formats for long-term preservation (Version 1.1). British Columbia, Canada: McLellan, E. P.

Jarvis, C.B., MacKenzie, S.B., & Podsakoff, P.M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*, 199-217.

Kalaian, S.A., & Kasin, R.M. (2012) Terminating sequential Delphi survey data collection. *Practical Assessment, Research & Evaluation, 17*(5).

Kongelige Bibliotek. (2004a, May). *Handling file formats*. Copenhagen: Denmark: Clausen, L.R.

Kongelige Bibliotek. (2004b, July). *Archival data format requirements*. Copenhagen: Denmark: Christensen, S.

Koninklijke Bibliotheek. (2008). *Evaluating file formats for long-term preservation*. The Hague, Netherlands: Rog, J., & Wijk, C, van.

Koninklijke Bibliotheek . (n.d.). *More about the e-Depot*. Retrieved June 7, 2013 from http://www.kb.nl/en/expertise/e-depot-and-digital-preservation/more-about-the-e-depot.

Lawrence, G.W., Kehoe, W.R., Rieger, O.Y., Walters, W.H., & Kenney, A.R. (2000). *Risk management of digital information: A File format investigation*. Washington, DC: Council on Library and Information Resources.

Lee, C. A. (2005). *Defining digital preservation work: A case study of the development of the reference model for an open archival information system*. Unpublished doctoral dissertation, University of Michigan.

Lesk, M. (1992). *Preservation of new technology: A report of the technology assessment advisory committee to the commission on preservation and access*. Washington, D.C.: Council on Library and Information Resources Commission on Preservation and Access.

Levy, D. (1998). Heroic measures: Reflections on the possibility and purpose of digital preservation. In *Proceedings of the third ACM conference on Digital Libraries*, New York, NY, 152-161.

Library of Congress. (n.d.). *The Archive Ingest and Handling Test (AIHT)*. Retrieved February 20, 2013 from http://www.digitalpreservation.gov/partners/aiht.html.

Luo, L., & Wildemuth, B. M. (2009). "Delphi studies." In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science* (chap. 10). Westport, CT: Libraries Unlimited.

Mace, G. M. & Lande, R. (1991, June). Assessing extinction threats: Toward a reevaluation of IUCN threatened species categories. *Conservation Biology, 5*, 148-157.

McEwen, S., & Goethals, A. (2009). File Information Tool Set (FITS): A new tool for digital preservation repositories. *D-Lib Magazine, 15*(9/10).

McGath, G. (2013). The format registry problem. *Code{4}Lib Journal, 19*. Retrieved January

24, 2014, from http://journal.code4lib.org/articles/8029.

Merriam-Webster. (1994). *Merriam-Webster's collegiate dictionary* (10th ed.). Springfield, MA: Merriam-Webster, Inc.

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis*. (2nd ed.). Thousand Oaks, CA: Sage Publications.

MITRE Corporation. (2009, December). *File format identification: Report on MITRE sponsored research*. Bedford, MA: Vidrine, K.

Moore, R. (2008). Towards a theory of digital preservation. *International Journal of Digital Curation 1*(3), 63-75.

National Digital Information Infrastructure & Preservation Program (NDIIPP). (2005, June). *Library of Congress Archive Ingest and Handling Test (AIHT): Final report*. Washington, DC: Shirky, C.

National Digital Information Infrastructure & Preservation Program. (2010). Preserving our digital heritage: The National Digital Information Infrastructure and Preservation Program 2010 report. Washington, D.C.

Naugler, H. (1984). *The archival appraisal of machine-readable records: A ramp study with guidelines*. Paris: General Information Programme and UNISIST United Nations Educational Scientific and Cultural Organization.

O'Grady, J.J., Reed, D.H., Brook, B.W., Frankham, R. (2004). What are the best correlates of predicted extinction risk? *Biological Conservation 118*, 513-520.

Oltmans, E., van Diessen, R.J., & van Wijngaarden, H. (2004). Preservation functionality in a digital archive. In *JCDL 2004: Proceedings of the Fourth Acm/Ieee Joint Conference on Digital Libraries: Global Reach and Diverse Impact: Tucson, Arizona, June 7-11, 2004*, edited by Hsinchun Chen, Michael Christel and Ee-Peng Lim, 279-86. New York, NY: ACM Press.

Online Computer Library Center [OCLC], Center for Research Libraries [CRL]. (2007). Trustworthy Repositories Audit and Certification (TRAC) criteria and checklist. Retrieved June 5, 2013 from www.crl.edu/PDF/trac.pdf

Online Computer Library Center [OCLC], Center for Research Libraries [CRL]. (2012) *ISO 16363:2012 Space data and information transfer systems -- Audit and certification of trustworthy digital repositories*. Retrieved December 27, 2013 from http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510

Open Planets Foundation. (2012). *Fido*. Retrieved May 11, 2013, from http://wiki.opf-labs.org/display/TR/Fido

Pearson, D., & Webb, C. (2008). Defining file format obsolescence: A risky journey. *International Journal of Digital Curation, 3*(1), 89-106.

Petter, S., Straub, D., & Rai, A. (2007). Specifying formative constructs in information systems research. *MIS Quarterly, 31*(4), 623-656.

Pontello, M. (n.d.).*TrID file identifier*. Retrieved February 21, 2013 from http://mark0.net/soft-trid-e.html

Public Health Agency of Canada. (2009). *Global Public Health Intelligence Network (GPHIN)*. Retrieved May 5, 2009 from http://www.phac-aspc.gc.ca/media/nr-rp/2004/2004_gphin-rmispbk-eng.php

Regents of the University of California. (2012). *Unified Digital Format Registry (UDFR): Final report.* San Francisco, CA.

Rieger, O.Y., and Kenney, A.R. (2000). *Risk management of digital information: Case study for image file format*. Washington, DC: Council on Library and Information Resources.

Rosenthal, D. S. H. (2010). Format obsolescence: assessing the threat and the defenses. *Library Hi Tech*, *28*(2), 195-210. doi:10.1108/07378831011047613

Ross, S. (2000). *Changing Trains at Wigan: Digital Preservation and the Future of Scholarship*. London: National Preservation Office.

Ross, S. (2007). Digital preservation, archival science and methodological foundations for digital libraries. Keynote address at the *11th European Conference on Research and Advanced Technology for Digital Libraries* (ECDL), Budapest (September 17, 2007).

Rothenberg, J. (1999a). *Avoiding technological quicksand: Finding a viable technical foundation for digital preservation*. Washington, D.C.: Council on Library and Information Resources.

Rothenberg, J. (1999b). *Ensuring the longevity of digital information*. Santa Monica, CA: Rand.

Rusbridge, C. (2006). Excuse Me... Some Digital Preservation Fallacies? *Ariadne*, *February 2*(46).

Scalable Preservation Environments (SCAPE). (n.a.). *About SCAPE*. Retrieved December 29, 2013 from http://www.scape-project.eu/about

Shaffer, M. L. (1981). Minimum for species population sizes conservation. *BioScience*, *31*(2), 131–134.

Soulé, M. E. (1985, December). What is conservation biology? *BioScience, 35*, 727-734.

Stanescu, A. (2004, November). Assessing the durability of formats in a digital preservation environment: The INFORM methodology. *D-Lib Magazine, 10*(11).

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology, 48*(5).

Sustainable Archives and Leveraging Technologies group (SALT). (2010). About CI-BER. Retrieved January 1, 2014 from http://ci-ber.blogspot.com/p/about-ci-ber.html

Task Force on Archiving of Digital Information. (1996). *Preserving digital information*. Washington, D.C.: Commission on Preservation and Access, Research Libraries Group.

Tessella. (2013). *Safety Deposit Box key features*. Retrieved February 21, 2013 from http://www.digital-preservation.com/solution/safety-deposit-box/sdb-key-features/

The National Archives. (2005a, July). *Selection of preservation formats: trends and issues*. Surrey, United Kingdom: Cornwell Management Consultants.

The National Archives. (2005b, July). *Criteria for the selection of preservation formats*. Surrey, United Kingdom: Cornwell Management Consultants.

The National Archives. (n.d.-a). *DROID: Digital Record Object Identification*. Accessed February 20, 2013 from http://freecode.com/projects/droid.

The National Archives. (n.d.-b). *The technical registry: PRONOM*. Retrieved June 7, 2013 from http://www.nationalarchives.gov.uk/PRONOM.

Thibodeau, K. (2002). *Overview of technological approaches to digital preservation and challenges in coming years: The state of digital preservation an international perspective*. Washington, D.C.: Council on Library and Information Resources.

Tibbo, H. R. (2003). On the nature and importance of archiving in the digital age. *Advances in Computers, 57*.

Underwood, W. (2009). *Extensions of the UNIX File command and magic file for file type identification* (Technical Report ITTL/CSITD 09-02). Atlanta, GA: Georgia Tech Research Institute.

Unified Digital Format Registry. (2011). *Unified Digital Format Registry (UDFR)*. Retrieved October 31, 2011 from http://www.udfr.org/

Unified Digital Format Registry Working Group. (2009a, March). *Proposal and roadmap.* Retrieved October 31, 2011 from http://www.gdfr.info/udfr_docs/Unified_Digital_Formats_Registry.pdf

Unified Digital Format Registry Working Group. (2009b, November). *Proposal to the National Digital Information Infrastructure and Preservation Program (NDIIPP) for technical development support for the Unified Digital Format Registry (UDFR)* (Draft, Version 4). Retrieved October 13, 2011 from http://www.udfr.org/docs/Udfr_proposal_nov2009_v4.doc

United Nations International Strategy for Disaster Reduction. (2006, October). *Global survey of early warning systems for Sustainable Development: International Strategy for Disaster Reduction*. Geneva.

van der Knijff, J. (2013a, September 30). *Assessing file format risks: searching for Bigfoot?* Message posted to Open Planets Foundation blogs at http://www.openplanetsfoundation.org/blogs/2013-09-30-assessing-file-format-risks-searching-bigfoot

van der Knijff, J. (2013b, October 8). *Measuring Bigfoot*. Message posted to Open Planets Foundation blogshttp://www.openplanetsfoundation.org/blogs/2013-10-08-measuring-bigfoot

Van De Ven, A. H., & Delbecq, A. L. (1974). The effectiveness of nominal, Delphi, and interacting group decision making processes. *The Academy of Management Journal*, *17*, 605-621.